13/11.

Volume 20, Part 1 *May 1967*

THE BRITISH JOURNAL OF MATHEMATICAL & STATISTICAL PSYCHOLOGY

EDITED BY R. J. AUDLEY

WITH THE ASSISTANCE OF CYRIL BURT

TAYLOR & FRANCIS LTD
RED LION COURT, FLEET STREET, LONDON, E.C.4

40*s.* *net* $7.45

Book Review Editor: A. R. JONCKHEERE

Abstracts Editor (Computer Programs): H. J. HALLWORTH

Editorial Board:

C. Banks
M. S. Bartlett
G. H. Bower
D. E. Broadbent
J. Brown
V. R. Cane
L. J. Cronbach
C. I. Howarth
A. R. Jonckheere
D. N. Lawley
R. D. Luce
S. Messick
E. A. Peel
L. S. Penrose
J. Fraser Roberts
W. Stephenson
A. Stuart
P. E. Vernon

PUBLICATIONS OF THE BRITISH PSYCHOLOGICAL SOCIETY

The British Journal of Mathematical and Statistical Psychology is issued by the British Psychological Society. Details of subscriptions to this and other Journals of the Society are given inside the back cover.

Papers for publication and books for review should be sent to the Editor, Professor R. J. Audley, Department of Psychology, University College London, Gower Street, W.C.1.

Abstracts of computer programs should be sent to Mr. H. J. Hallworth, Department of Education, University of Birmingham, Birmingham 15. Articles on computers, computer languages and programming will also occasionally be published and should be addressed to Mr. Hallworth.

PREPARATION OF PAPERS

Contributors are asked to send their papers in a form which is ready for submission to the printer: that is to say, the arrangement of the material should follow that observed in a current issue of this *Journal*. Each article should be headed with a brief abstract. Each section should have a short cross-heading, in capitals; and the whole should conclude with a summary embodying the main conclusions reached. Special attention should be paid to such details as correct spelling, grammar, capitalization, numbering, etc. In preparing manuscripts and in correcting proofs authors should conform, so far as possible, with the 'Recommendations to Authors' set out in *The Printing of Mathematics* by T. W. Chaundy, P. R. Barrett, and Charles Batey (Oxford University Press, 1954, ch. 2, pp. 21–73), or to those given in the paper on 'Setting Mathematics' (Arthur Phillips, *Monotype Recorder*, 60, iv, 1956).

Owing to the present conditions for printing, the cost of setting up tables and algebraic formulae which require hand composition is very much greater than the cost of machine composition. Contributors are therefore advised to arrange their material so that it can be printed as economically as possible. Often numerical results and algebraic formulae can, with a little simplification, be readily adapted for purposes of machine composition. Authors must ensure that formulae are clear, particularly when these contain subscripts.

Tables, photographs and figures which are costly to print should be used only when they save considerable space or are essential for effective presentation; and in no case should both a figure and table be used to convey the same information. Each table or figure should be numbered and on a separate sheet with its approximate position indicated in the typescript.

Drawings should be in Indian ink on heavy unruled paper. Lettering on graphs, etc. is best indicated in a tracing placed over the diagram. Diagrams are normally reduced in reproduction and allowance must be made for a corresponding reduction in the thickness of lines and the spaces between them. The numbered legend of a figure should be typewritten on a separate sheet of paper, not lettered on the drawing.

The object of the caption for a table or the legend for a figure should be to make the table or the figure comprehensible without reference to the text.

References in the text should be indicated by dates in parentheses after the authors' names and correspond to an alphabetical list at the end of the paper. The list must follow the layout illustrated in the three following examples:

ALLPORT, G. W. & VERNON, P. E. (1931). A test for personal values. *J. abnorm. soc. Psychol.* **26**, 231–48.

GALTON, F. (1883). *Inquiries into Human Faculty and its Development.* London: Macmillan.

REVESZ, G. (1951). Le probleme du genie. *L'Annee psychol.* **50**, 83–96.

World list abbreviations are used for journals. The Oxford Dictionary is taken as the authority on spelling. For other general guidance, authors are referred to *The Journal of Physiology: Suggestions to Authors*. Cambridge University Press. Price 2*s*.

In future authors will receive one complete slip-proof of their papers, but no page proof. They will be expected to pay the cost of all corrections for which they themselves are responsible.

All manuscripts should be submitted in duplicate.

PROGRAM ABSTRACTS

An abstract of a computer program, normally of 300 to 500 words in length, should state the sequence of standard mathematical or statistical operations that are performed, the language in which the program is written, the approximate running time and should either give an address from which the program can be obtained (in tape or card form) or state where data may be sent for processing. Authors are referred to a current issue of the *Journal* for examples of an acceptable abstract. In most cases, the abstract should be accompanied by an actual program in the form of a deck of cards or a paper tape together with the appropriate listing information, also any other instructions which are necessary for the preparation of data or the running of the program.

Vol. 20
Part 1
1–29

The British Journal of Mathematical and
Statistical Psychology

May
1967

32

REACTION-TIME STUDIES: THE ANTICIPATION AND INTERACTION OF RESPONSES

By E. A. C. Thomas
St. John's College, Cambridge

BUREAU OF EDUCATION L AND
GOVT. OF W.
LIBRA
CALCUTTA

The present study is largely concerned with the relation between reaction time and the temporal uncertainty of the signal. It is suggested that the appropriate measure of signal uncertainty is the conditional probability that the signal would be presented given that it has not been presented before. This conditional probability is called the expectancy function. It is assumed that the expectancy function influences the subject's state of readiness and this in turn influences reaction time; and ways of expressing these relations mathematically are discussed. It is shown that, with a suitable representation of subjective errors in time and probability estimation, a theory can be stated which accounts for the qualitative features of a variety of experimental data.

The relevance of this theory to studies concerned with central intermittency is then discussed. It is argued that temporal uncertainty alone is insufficient to account for the increase in reaction time to the second of two closely spaced signals, both of which require a response. On the other hand, a model for the interaction of two responses is stated in which refractoriness in the central processes is not inevitable.

BUREAU OF EDUCATIONAL AND PSYCHOLOGICAL RESEARCH
GOVT. OF W. BENGAL
LIBRARY
CALCUTTA-19

1. Introduction

With the growing number of attempts by psychologists to construct a quantitative treatment of sensori-motor performance has developed a revival of interest in reaction-time studies. In these experiments the subject is told to expect a signal (usually a flash of light or a pure tone well above threshold intensity), and to respond (usually by pressing a key) as quickly as possible after perceiving the signal. The reaction time is the time between the presentation of the signal and the making of the response. At every trial the subject is given a warning signal, and after a variable foreperiod the signal is presented.

There is a classification of reaction-time experiments based on the number of alternatives from which the signal may be chosen. In the simple reaction-time task the subject knows what the signal is going to be and his only uncertainty is as to when it will be presented. In the disjunctive situation the signal can be one of a set of alternatives, thus adding another source of uncertainty. Further, more than one signal of the set may correspond to the same response or more than one response to the same signal. In the serial reaction task two signals are presented in succession at each trial and the subject has to respond to both. Here both the foreperiod and the inter-signal interval are variable but are small compared with the inter-trial interval. Studies in this group have been concerned mainly with central intermittency, and their relationship to those on

Bureau of Ednl & Psyl Research
(S. C. E. R. T.)
Date 7.8.95

simple and disjunctive reaction times seems far from clear. For a review of some of the issues involved the reader is referred to the article by Bertelson (1966).

The main aim of this paper is to discuss possible mathematical models of the relation between subjective readiness and simple reaction time. Studies on central intermittency will then be discussed in the light of these models. The next section is a review of some of the large body of experimental data which provided a starting point for the present study; and in later sections more data will be introduced to illustrate specific points. The following abbreviations will be used: S refers to the subject, Σ to the signal, R to the response, RT to the average reaction time and SR to the subject's state of readiness. These abbreviations will sometimes appear with suffixes, but the meanings should be clear.

2. Some Experimental Results

In all the experiments reported below, the readings used were taken after Ss had had extensive practice in the particular condition. Klemmer (1956) investigated the relation between foreperiod variability and RT by using two kinds of tests: Interval and Foreperiod tests. In the first kind of test a typical series consisted of Σ being presented at regular intervals of length τ, which differed among tests. These tests were later repeated by Karlin (1959) and, within the range 0·5 sec to 12 sec, it was found that RT increased with τ and the slope of the curve decreased with τ, i.e. the curve was concave.

In a typical series of the second kind, the foreperiod at a trial was drawn from an eleven-point rectangular frequency distribution on $(s, s+h)$. s, the minimum foreperiod, and h, the foreperiod range, were different for different series. The two statistics of interest were the RT to a signal when it came after a particular interval, and the mean RT to all signals in a series. It was found that the former RT varied little with the interval, but that the latter was an increasing function of h, though it showed little change when s was varied from 1·25 to 7·25 sec, h being held constant. Klemmer concluded that 'S's time uncertainty is the result of both his own imperfect time-keeping ability and the clock-time variability of the stimulus'. The first factor is perhaps reflected by the RT-τ curves, and the second by the dependence on h of the mean RT for a series.

To estimate the effects of each source of uncertainty two more kinds of experimental series were used (Klemmer, 1957). The first kind was the same as the Foreperiod tests mentioned above, except that a normal frequency distribution with mean μ and standard deviation (s.d.) σ was used. To each series of this kind there was a corresponding series of the second kind, the Prediction tests, in which S was asked at each trial to estimate an interval of length μ. S received immediate knowledge of the magnitude and direction of his error of prediction. The variance of the prediction times was taken as a measure of S's uncertainty in estimating an interval of length μ. The total time uncertainty, σ_T^2, for the particular Foreperiod series (μ, σ) was found by adding this prediction variance to σ^2. Klemmer found that RT was linearly related to $\log \sigma_T$. It should be noted that if there is no subjective time uncertainty, then

$$\begin{aligned}\log \sigma_T &= \log \sigma \\ &= H - \text{constant},\end{aligned}$$

where H is the Shannon information measure for the normal distribution. So, from the result that RT is linearly related to H (Hick, 1952), one might regard $\log \sigma_T$ as the ' information ' to the subject whose time-keeping is imperfect.

The four values of μ used in the Prediction tests were 2^{n-1} sec, $n=0$, 1, 2 and 3. Let the observed s.d. be denoted by σ_n. An inspection of the results for the 5 *S*s (Klemmer, 1957, Figure 1) shows that, at least approximately,

$$\begin{aligned}\log_2 \sigma_n - \log_2 \sigma_{n-1} &= c, \text{ a constant.} \\ \therefore \quad \sigma_n &= 2^c \sigma_{n-1} \\ &= (2^c)^n \sigma_0 = (2^c)^{n-1} k_0 \\ &= (2^{n-1})^c k_0.\end{aligned}$$

Therefore, the s.d., $\sigma(\tau)$, of *S*'s estimation of an interval of length τ is approximately given by

$$\sigma(\tau) = k_0 \tau^c , \tag{2.1}$$

where k_0 and c are constants. For one S, $c \doteq 1{\cdot}1$, and for the other four *S*s, $c \doteq 0{\cdot}8$.

More recent work on Foreperiod tests has revealed a significant relation between RT and its immediate foreperiod (Davis, 1959; Karlin, 1959; Drazin, 1961). Davis found that RT decreases linearly as immediate foreperiod increases from 0·05 sec to about 0·25 sec, after which RT remains constant. Drazin used a continuous rectangular foreperiod distribution and has shown that, generally, the RT-foreperiod curve is decreasing with decreasing slope. It was also shown that the RT at the minimum foreperiod increases as s decreases from 2 sec to 0·125 sec. This suggests that there is a critical value of s below which S cannot make full use of the warning signal (see also Klemmer, 1956). Deviations from linearity at the smallest foreperiods are also evident in Davis (1959).

In another experiment (Mowbray, 1960), which falls somewhere between the Foreperiod and Interval tests, S was required to respond only to a particular signal out of a set of n signals. At each trial each signal was equally likely to be presented, the presentation rate being fixed at one signal every 2·5 sec. It was found that RT increased linearly with n in the range $2 \leqslant n \leqslant 10$.

In some experimental series there was a fixed proportion of ' catch ' trials in which no signal appeared after the warning signal (Davis, 1956; Drazin, 1961). One reason for including them was to check whether S had a tendency to anticipate his reaction. Whenever, and only when, this was done, there was a tendency for the minimum RT to occur at an interval less than the maximum. The other comparable experiment in which this effect appeared was that of Elithorn and Lawrence (1955), and in this case there were trials on which only one signal was presented though S was warned to expect the omission. Mowrer (1940), using a different type of ' catch ' trial, observed an effect to which the above effect has been compared, somewhat misleadingly. Mowrer trained his Ss to expect a signal after a fixed interval and then presented them with catch trials at which the signal came at intervals different from the fixed interval. He found that the RT for the catch trials was greater than the RT for the fixed interval trials. On the basis of this finding, Davis (1956) has suggested that ' where several intervals are used in the training situation the subject reacts most quickly

(*a*) at the most frequently presented interval,

(*b*) at the interval in the middle of the range.'

Still more recent work (Davis, 1965) has attempted to show that the relation between RT and foreperiod is insignificant. In this experiment the warning signal at each trial was a response made by S whenever he chose. Karlin (1965) has already pointed out that this experiment is inconclusive because, in subjective terms, the interval between an *ad lib.* response and a subsequent signal may not be equivalent to the interval between two signals. This argument should become clearer later on. Other work (Bertelson and Boons, 1960; Kay and Weiss, 1961; Borger, 1963), though not concerned specifically with the RT-foreperiod relation, has tended to be consistent with the above-mentioned work and will be referred to later on in the paper.

3. RT, SR AND THE EXPECTANCY FUNCTION

The results quoted above suggest that when S knows what signal to expect but not when it will arrive, there are at least two relevant processes occurring in time. The first describes the time-course of S's state of readiness (SR) up to the time that Σ arrives, and is perhaps related to the objective stimulus conditions. The second process, dependent on the first, starts when Σ arrives and ends with the making of a response, and involves the recognition of Σ as the signal requiring a response and the organization of the response. For the present, no attempt will be made to specify the way in which this second process evolves, and attention will be confined to the possible relation of its length, i.e. RT, to the endpoint of the first process.

One of the first requirements of this approach is to specify the independent variable of which SR is a function. This variable would itself be a function of time (here assumed discrete) and the experimental conditions and, at every time t, should measure the signal uncertainty at t conditional on the process up to t. A natural measure of this sort is p_t, the conditional probability that Σ would arrive at t given that it has not arrived before t, and this is the measure which will be *assumed* to be relevant. It will be called the *expectancy function*.

The next step is to characterize SR and then to make assumptions about its dependence on p_t. Two ways of specifying SR might be by measuring levels of muscular tension or by measuring the electrical activity in the cortex. As far as the author is aware, more attention has been paid recently to the cortical correlates of stimulus uncertainty than to the muscular. For this reason and because these two measurements are probably highly correlated, we shall consider the experimental findings concerning only cortical activity. In their study of waveforms evoked by auditory and visual stimuli, Walter *et al.* (1964) have shown that expectancy can correspond to the development of a large negative component in the waveform; this component is called the contingent negative variation (*CNV*). When *S* is told that a warning signal would be followed by another signal requiring a response, a large *CNV* appears rising to a peak just after the second signal is presented and then ending abruptly. They found that the introduction of catch trials tended to reduce the *CNV* (mixing 27 catch trials randomly in a series of 48 reduced the *CNV* by about a half), though for most *S*s the *CNV* showed signs of attenuation only when the proportion of catch trials exceeded a critical value (about 0·3). It was also found that in most *S*s the development of a consistent *CNV* with an abrupt ending was associated with a marked reduction in *RT* to the signal.

These results suggest, firstly, that it is possible to characterize *SR* such that it increases as p_t increases up to some value, p_x, beyond which it is constant. The weak assumption will be made that *SR* is measured in units which make it directly proportional to p_t for $p_t \leqslant p_x$. To consider the effect on *SR* of time-keeping errors it will be assumed that *S*'s error in estimating a time-point at which Σ may arrive is small, and that his counting off of successive time-points is perfect. The assumption of small time-keeping errors seems reasonable in view of the results of Woodrow (1930) showing that subjective estimates of time intervals in the range 0·2 sec to 1·5 sec had a standard deviation which was approximately 10 per cent of the standard interval. These assumptions enable us to proceed after specifying the *SR* only in a small, continuous neighbourhood of the time at which *S expects* Σ. Thus if *T* is *S*'s estimate of *t*, then, in the absence of a signal, *SR* would rise to a local maximum proportional to p_t at *T* and then decline.

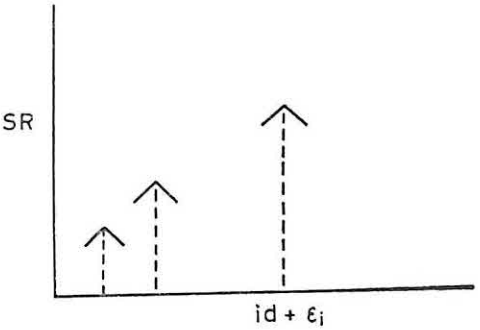


FIGURE 1. Time-course of subjective readiness (*SR*) when foreperiod distribution is discrete; ϵ_i is *S*'s error in estimating the *i*th time-point.

Let it be assumed that, for z in a continuous neighbourhood of *T*, *SR* has the expansion,

$$SR(z) = lp_t - m|z-T| + m'|z-T|^2 + \dots,$$

If $|z-T|$ is small, then, as an approximation,

$$SR(z)=lp_t-m|z-T|\,, \tag{3.1}$$

where l and m are positive constants. This situation is shown in Figure 1.

Secondly, we may suppose that RT decreases as SR increases. So the RT to Σ which comes at t is given by

$$\begin{aligned} RT_t &= f[SR(t)] \qquad (0<p_t\leqslant p_x)\ ; \\ &= RT_m \qquad (p_t\geqslant p_x)\ ; \end{aligned} \tag{3.2}$$

where $f(.)$ is a decreasing function of its argument and RT_m denotes the minimum RT. Three simple forms of $f(.)$ are the following:

$$\begin{aligned} f(x) &= a+\frac{b}{x}\ ; && \text{(a)} \\ &= a-bx\ ; && \text{(b)} \\ &= a-b\log x. && \text{(c)} \end{aligned} \tag{3.3}$$

There is little evidence to help in the choice of $f(.)$, but it will now be shown that the RT curves given by eqn. (3.2), where $f(.)$ is given by (3.3 a), give a good fit to the data of Davis (1957, 1959) and Mowbray (1960).

In the experiments of Davis the foreperiod distribution was uniform on the integers 1 to n. Thus, by Bayes' rule,

$$\begin{aligned} p_t &= \frac{1/n}{1-(t-1)/n} \\ &= (n-t+1)^{-1}, \qquad t=1,2,\ldots,n. \end{aligned} \tag{3.4}$$

If $f(.)$ is given by eqn. (3.3 a), then for the ideal S whose timekeeping is perfect, it follows from (3.1) and (3.2) that

$$\begin{aligned} RT_t &= a'-b't \qquad 1\leqslant t\leqslant j \\ &= RT_m \qquad j\leqslant t\leqslant n \end{aligned}$$

for some j. This linear relation is in accord with the experimental results. In the experiment of Mowbray there were n possible signals and the probability that a particular one would be presented at a given trial was, therefore, $1/n$. So that, for the ideal S, if (3.3 a) holds, the RT to the particular Σ when there are n possible ones would be given by

$$RT(n)=a''+b''n.$$

This linear increase of $RT(n)$ with n was in fact observed.

Thus it is concluded that the expression (3.3 a) for $f(.)$ is not unreasonable. Further tests of its adequacy can be carried out by comparing observed and expected RT curves using different foreperiod distributions. However, (3.3 a) seems simpler to handle mathematically and will be adopted here mainly to provide a theoretical framework within which certain ideas on expectancy can be discussed. The discussion will be limited to two qualitative aspects of *RT*-foreperiod curves when the foreperiod distribution is uniform: the sign of the slope and that of the rate of change of the slope of these curves. Following a convention, we shall say that a curve is convex (concave) when this latter sign is positive (negative).

First to be discussed are the Interval tests where there is no temporal uncertainty, so that *S*'s only error is in estimating the interval between signals. It is assumed that *S*'s estimate of a time interval is equal to the true interval length plus an error term. This linearity assumption about subjective time estimates is retained in the discussion of the discrete Foreperiod tests. It is shown that this way of representing time-keeping errors provides a means of accounting for certain features of empirical *RT* curves. In the treatment of the continuous Foreperiod tests no explicit mention of time-keeping errors will be made and it is assumed only that *S* adjusts his *SR* intermittently. In Section 5 an alternative mathematical treatment of the problem will be considered.

4. MATHEMATICAL MODELS

4.1 *The Interval tests*

In these experiments Σ is presented after a fixed foreperiod, τ, and the effect on $RT(\tau)$ of *S*'s time-keeping imperfection has to be considered. It is supposed, following eqn. (3.1), that *SR* rises linearly with gradient m so as to reach a maximum, l, at the point in time, T, at which *S* expects Σ. Here, however, it seems reasonable to suppose that if Σ does not arrive by T, this maximum is maintained until Σ does arrive (see Figure 2). If we write $T = \tau + \epsilon$, then ϵ is the prediction error studied by Klemmer (1957).

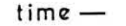


FIGURE 2. Time-course of *SR* when foreperiod is a constant, τ.

S's *RT* would be lengthened whenever $\epsilon > 0$, because in this case the maximum *SR* would not have been attained by the time τ. Let E denote 'mathematical expectation'. Then from eqns. (3.1), (3.2) and (3.3 a) it follows that $RT(\tau)$ is proportional to

$$\frac{1}{l}\, p(\epsilon \leqslant 0) + E\left[\frac{1}{l - m\epsilon} \,|\, \epsilon > 0\right] + \text{constant},$$

If it is assumed that ϵ is normally distributed with mean 0 and s.d. σ, then

$$p(\epsilon \leqslant 0) = 1/2,$$

and the probability that $x \leqslant \epsilon \leqslant x + dx$ given that $\epsilon \geqslant 0$ is

$$\frac{2}{\sigma\sqrt{2\pi}} \exp\left(-x^2/2\sigma^2\right) dx, \qquad x \geqslant 0.$$

$$\begin{aligned}\therefore \quad E(\epsilon|\epsilon \geqslant 0) &= \frac{1}{\sigma}\sqrt{\left(\frac{2}{\pi}\right)}\int_0^\infty x \exp\left(-x^2/2\sigma^2\right) dx \\ &= \sigma\sqrt{\frac{2}{\pi}}.\end{aligned}$$

So that, neglecting terms of order ϵ^2 and higher,

$$\begin{aligned} E\left[\frac{1}{l - m\epsilon}|\epsilon \geqslant 0\right] &\doteq \frac{1}{l} E\left[1 + \frac{m\epsilon}{l}|\epsilon \geqslant 0\right] \\ &= \frac{1}{l}\left[1 + \frac{m\sigma}{l}\sqrt{\left(\frac{2}{\pi}\right)}\right].\end{aligned}$$

$\therefore \quad RT(\tau) = a_0 + a_1\,\sigma(\tau)$, where a_0, $a_1 > 0$ and, from eqn. (2.1), $\sigma(\tau) = k_0\tau^c$.

$$\therefore \quad \frac{d}{d\tau} RT(\tau) = ca_1k_0\tau^{c-1} > 0,$$

and

$$\frac{d^2}{d\tau^2} RT(\tau) = -c(1-c)a_1k_0\tau^{c-2}.$$

Hence if $0 < c < 1$, an assumption suggested by the results of Klemmer (1957), the second derivative is negative. In other words, for small τ, the curve $RT(\tau)$ is a concave, increasing function of τ. For larger values of τ the term containing ϵ^2 needs to be considered. Then, since $E(\epsilon^2|\epsilon \geqslant 0) = \sigma^2$,

$$\frac{d^2}{d\tau^2} RT(\tau) = A\tau^{c-2}[\tau^c - B(1-c)],$$

where A and B are positive constants, which is negative for $\tau < [B(1-c)]^{1/c}$. This result seems to be consistent with the curves obtained by Klemmer (1956) and Karlin (1959).

4.2. *The Foreperiod tests, discrete case*

It is supposed that in these experiments Σ arrives at clock-time td, where d (secs) is a constant and t is chosen at random from the integers $1, 2, \ldots, n$. Then as pointed out in Section 3, the RT curve is the following:

$$\begin{aligned} RT_t &= a' - b't & 1 \leqslant t \leqslant j \\ &= RT_m & j \leqslant t \leqslant n \end{aligned}$$

for some j.

S's time-keeping imperfection can be accounted for in the following manner. Suppose that Σ arrives at time id; then S has had to 'predict' each of the time-points td, $t=1, 2, \ldots, i$. It is assumed that S predicts one time-point starting from the previous one, so that for each prediction S predicts an interval of length d and does so with an error ϵ. It is also assumed that ϵ is $N(0, \sigma^2)$, where $\sigma \ll d$. Then the error, ϵ_i, in predicting the interval of length id is $N(0, i\sigma^2)$. The SR curve for this situation is shown in Figure 1.

$$\begin{aligned} RT_i &= a+bE\left[\frac{1}{p_i-m|\epsilon_i|}\right] \doteq a+\frac{b}{p_i-mE(|\epsilon_i|)} \\ &= a+b[(n-i+1)^{-1}-g\sqrt{i}]^{-1}, \quad \text{where } g\left[=m\sigma\sqrt{\left(\frac{2}{\pi}\right)}\right] \ll 1, \\ &\doteq a+b(n-i+1)+bg(n-i+1)^2\sqrt{i}, \end{aligned} \tag{4.2.1}$$

neglecting terms of order g^2 and higher.

The following restriction on g is imposed:

$$(n-i+1)^{-1}-g\sqrt{i}>0 \text{ for } i=1, 2, \ldots, n.$$

This is satisfied if

$$g<\frac{3\sqrt{3}}{2}(n+1)^{-3/2}. \tag{4.2.2}$$

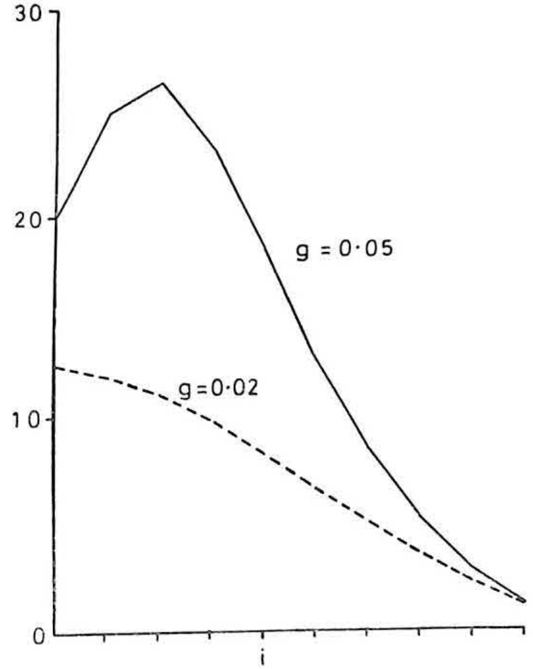


FIGURE 3. RT-foreperiod curves. Plot of $[(n-i+1)^{-1}-g\sqrt{i}]^{-1}$ against i, for $n=10$, $g=0{\cdot}05$, $0{\cdot}02$.

Treating i as continuous, it follows from eqn. (4.2.1) that

$$\frac{d^2RT_i}{di^2} = \frac{bgi^{-3/2}}{4}[15i^2 - 6(n+1)i - (n+1)^2],$$

which is negative for $i < 7{\cdot}9(n+1)/15$ approximately, and positive for the rest of the range. In other words, the curve is concave for about the first half of the range and convex for the other.

It is easy to show that (4.2.2) implies that $RT_n < RT_{n-1}$. However, $RT_2 > RT_1$ if

$$(n-1)^{-1} - g\sqrt{2} < n^{-1} - g,$$

i.e. if

$$g > [n(n-1)(\sqrt{(2)} - 1)]^{-1}. \tag{4.2.3}$$

It can be seen that (4.2.2) and (4.2.3) are compatible only if $n > 4$. So that if $n > 4$ it is possible that the maximum RT occurs not at the shortest foreperiod but at a longer one (see Figure 3). An effect of this kind has been observed by Elithorn and Lawrence (1955) and by Klemmer (1956).

4.3. *The inclusion of ' catch ' trials*

Consideration will now be given to a foreperiod series with a uniform frequency distribution of foreperiod in which a proportion $1 - \pi$ of catch trials is included. Then

$$\begin{aligned} p_t &= \frac{\text{prob. } (\Sigma \text{ arrives at } t)}{\text{prob. } (\Sigma \text{ arrives at } t \text{ or after})} \\ &= \frac{\pi/n}{1 - \pi(t-1)/n} = \pi[n - \pi(t-1)]^{-1}. \end{aligned} \tag{4.3.1}$$

So that for the ideal S, for some j,

$$\begin{aligned} RT_t &= a''' - b'''t \qquad 1 \leqslant t \leqslant j \\ &= RT_m \qquad j \leqslant t \leqslant n. \end{aligned}$$

This linear decrease is contrary to the experimental results of Davis (1956) and Drazin (1961), who found that RT_t reached a minimum for some $t < n$ and then increased for larger t. The introduction of time-keeping errors would not account for this feature since, as was shown in the previous section, RT_n is less than RT_{n-1}, provided the errors are small and $n > 4$.

In deriving eqn. (4.3.1) it was assumed that the subjective belief that a trial is not a catch trial is unchanged while S waits for Σ. However, it is likely that this subjective belief decreases while S waits for Σ and this idea is now developed.

The π in (4.3.1) is replaced by π_t, the subjective probability at time t that the trial is not a catch trial; and the substitution $\omega_t = \pi_t^{-1}$ is made. The turning point (minimum) of p_t^{-1} is required. From (4.3.1),

$$\begin{aligned} p_t^{-1} - p_{t-1}^{-1} &= n(\omega_t - \omega_{t-1}) - 1 \\ &= 0 \quad \text{when} \\ \omega_t - \omega_{t-1} &= 1/n. \end{aligned} \tag{4.3.2}$$

It remains to make assumptions about the form of ω_t. The following three assumptions seem acceptable on intuitive grounds:

(i) $\omega_1 = \pi^{-1}$, $\omega_t - \omega_{t-1} \geqslant 0$, i.e. S's belief at the start of a trial that Σ would be presented matches the objective probability, and then declines.

(ii) $\omega_t = 1$ for all t when $\pi = 1$, i.e. if no catch trials are included, S does not waver in his belief that Σ would be presented while waiting for it.

(iii) The rate at which S loses his belief is small initially and then increases, i.e. π_t is concave. This implies that ω_t is convex.

A fourth assumption is suggested by the empirical finding that $RT_1 > RT_n$, which implies that $p_1^{-1} > p_n^{-1}$, and hence

(iv) $\omega_n < \pi^{-1} + (1 - 1/n)$.

That these four assumptions do not guarantee that (4.3.2) will always have a solution $\theta\,(<n)$ can be seen by putting

$$\omega_n = \pi^{-1} + (1 - 1/n) - \delta, \qquad 0 < \delta < (1 - 1/n)$$
$$\omega_t - \omega_{t-1} = n^{-1} - \delta/(n-1).$$

For then all four conditions are satisfied, but

$$\omega_t - \omega_{t-1} < 1/n \text{ for } 1 \leqslant t \leqslant n.$$

On the other hand, if (4.3.2) has a solution $\theta < n$, then

$$\omega_t - \omega_{t-1} > 1/n \text{ for all } t > \theta.$$

Or, to put it crudely, for (4.3.2) to have a solution the convexity of ω_t must not be too small. The two types of expressions for ω_t that will be considered are: (*a*) a polynomial of the form

$$\omega_t = \sum_{j=0}^{k} c_j t\,, \tag{4.3.3}$$

where the c's are positive; and
(*b*) the exponential expression

$$\omega_t = d_0 + d_1 \pi^{-t}, \qquad d_1 > 0. \tag{4.3.4}$$

Eqn. (4.3.3) satisfies conditions (i) to (iii) if the c's are given by

$$c_0 = \pi^{\alpha_k},$$
$$c_j = \pi^{\alpha_{\;-1}} - \pi^{\alpha_j}, \qquad j = 1, 2, \ldots, k,$$

where $-1 = \alpha_0 \leqslant \alpha_1 \leqslant \ldots \leqslant \alpha_k$. For simplicity it will be assumed that $-1 = \alpha_0 = \alpha_1 = \ldots = \alpha_{k-1} < \alpha_k = \alpha$. Eqn. (4.3.4) also satisfies these three conditions provided

$$\pi d_0 = 1 - d_1.$$

It will now be shown that solutions, θ, of eqn. (4.3.2) such that $\theta < n$ are possible, but that all these solutions exceed $(n+1)/2$. The following results will be used:

$$n^{k+1} - (n-1)^{k+1} > 2n^k - 1\,; \tag{4.3.5}$$
$$(n+1)^k - (n-1)^k < 2^k(n^k - 1)(n-1)^{-1}; \tag{4.3.6}$$

for $n > 3$, $k \geqslant 2$.

The sequences $\{u_n\}$ and $\{v_n\}$ are decreasing, where

$$u_n = n\pi^{n-1} - (n-1)\pi^n, \tag{4.3.7}$$

and

$$v_n = (1-\pi^n)/n; \tag{4.3.8}$$

for $\pi < 1$.

The inequalities (4.3.5) and (4.3.6) can be established by induction on k. For, putting $k=2$ in (4.3.5), it follows that

$$n^3 - (n-1)^3 = 2n^2 - 1 + (n-1)(n-2) > 2n^2 - 1.$$

If (4.3.5) holds for k, then

$$\begin{aligned} n^{k+2} - (n-1)^{k+2} &= nn^{k+1} - (n-1)^{k+2} \\ &> n[2n^k + (n-1)^{k+1} - 1] - (n-1)^{k+2} \\ &= 2n^{k+1} - 1 + (n-1)[(n-1)^k - 1] \\ &> 2n^{k+1} - 1, \end{aligned}$$

so that it holds for $k+1$ and, therefore, for all k. The proofs of (4.3.6) to (4.3.8) are omitted.

A theorem about the ω_t will now be proved.

THEOREM 4.3

If ω_t, given by eqn. (4.3.3) or (4.3.4), satisfies conditions (i) to (iv), then
(*a*) solutions, θ, of (4.3.2) such that $1 < \theta < n$ are possible, and
(*b*) if a solution exists, then $\theta > (n+1)/2$.

PROOF

Case (1). $\omega_t = \pi^\alpha + (\pi^{-1} - \pi^\alpha)t^k, \quad k \geqslant 2.$

Condition (iv) implies that

$$[n(\pi^{-1} - \pi^\alpha)]^{-1} > (n^k - 1)(n-1)^{-1}.$$

From (4.3.2),

$$[n(\pi^{-1} - \pi^\alpha)]^{-1} = \theta^k - (\theta - 1)^k,$$

from which it can be seen that θ cannot be less than n unless

$$n^k - (n-1)^k > (n^k - 1)(n-1)^{-1}, \quad \text{or}$$
$$n^{k+1} - (n-1)^{k+1} > 2n^k - 1,$$

which from (4.3.5) is true. So that a solution $\theta < n$ is possible.

To show that, if a solution exists, $\theta > (n+1)/2$, it is necessary only to show that

$$\left(\frac{n+1}{2}\right)^k - \left(\frac{n-1}{2}\right)^k < (n^k - 1)(n-1)^{-1},$$

which is the inequality (4.3.6).

Case (2). $\omega_t = (1-d)\pi^{-1} + d\pi^{-t}.$

Condition (iv) implies that

$$d < \pi^n(n-1)[n(1-\pi^{n-1})]^{-1}.$$

From (4.3.2),

$$d = \pi^{\theta}[n(1-\pi)]^{-1},$$

so that $1 < \theta < n$ only if

$$\pi^{n}[n(1-\pi)]^{-1} < \pi^{n}(n-1)[n(1-\pi^{n-1})]^{-1} < \pi[n(1-\pi)]^{-1}.$$

The left-hand inequality reduces to

$$1-\pi > (1-\pi^{n-1})(n-1)^{-1}, \quad \text{or}$$
$$v_1 > v_{n-1}, \text{ which follows from (4.3.8).}$$

The right-hand inequality reduces to

$$(n-1)(1-\pi)\pi^{n-1} < 1-\pi^{n-1}, \quad \text{or}$$
$$n\pi^{n-1} - (n-1)\pi^{n} < 1, \quad \text{or}$$
$$u_n < u_1, \text{ which follows from (4.3.7).}$$

Therefore a solution $\theta < n$ is possible.

If a solution exists then

$$\pi^{\theta}[n(1-\pi)]^{-1} < \pi^{n}(n-1)[n(1-\pi^{n-1})]^{-1}, \quad \text{or}$$
$$\pi^{n-\theta} > (1-\pi^{n-1})[(n-1)(1-\pi)]^{-1}.$$

To show that $\theta > (n+1)/2$ it has to be shown that

$$\pi^{(n-1)/2} < (1-\pi^{n-1})[n-1)(1-\pi)]^{-1}, \quad \text{or}$$
$$y(\pi) \equiv 1-\pi < (\pi^{-m} - \pi^{m})/2m \equiv z(\pi),$$

where $m = (n-1)/2 > 1$.

But

$$\left.\frac{dz}{d\pi}\right|_{\pi=1} = -1 = \left.\frac{dy}{d\pi}\right|_{\pi=1},$$

and

$$\frac{d^2z}{d\pi^2} = \frac{1}{2\pi^{m+2}}[m(1-\pi^{2m}) + 1 + \pi^{2m}] > 0 = \frac{d^2y}{d\pi^2}.$$

$\therefore\ z(\pi) > y(\pi)$ for $0 < \pi < 1$,

which completes the proof of the theorem.

The above theorem gains some generality if one accepts that any convex, increasing function, ω_t, can be fairly well approximated in the range $1 \leqslant t \leqslant n$ by a curve either of the form (4.3.3) or of the form (4.3.4). The theorem then leads us to expect that when catch trials are included, the minimum RT would occur at a point greater than the midpoint of the range. This result seems to be at variance with the suggestion (Davis, 1956) that maximum preparedness, and hence minimum RT, occurs at about the midpoint of the range.

θ can be obtained from the experimental data, and (4.3.2) can then be solved for α or d. It would then be possible, using this estimate of α or d, to estimate the value of $t(<\theta)$ for which RT_t is approximately equal to RT_n. This value of t can then be compared with the observed value. The parameters α

and d are thought of as being possibly dependent on S but independent of π. These predictions, together with that of Theorem 3.4, can be regarded as tests of the adequacy of the above approach.

4.4 *The Foreperiod tests, continuous case*

In Section 4.2 it was implicit that S was aware of the discrete nature of the foreperiod distribution so that he was able to build up his SR to reach local maxima at the distribution points. When the foreperiod distribution is subjectively continuous (assumed uniform on the interval $(s, s+h)$), there are no such preferred time-points at which SR should be maximized. However, one can assume that, as in the discrete case, SR is adjusted at discrete points, but, since the distribution is now continuous, it is necessary to specify SR for all t, $s \leqslant t \leqslant s+h$. Let it be assumed that S adjusts his SR at instants s, $s+w_1$, $s+w_1+w_2, \ldots, s+w_1+\ldots+w_N=s+h$, and that in the nth interval of length w_n his SR is proportional to the conditional probability, p_n that Σ would arrive in that interval. Then

$$p_n = w_n/(w_n+w_{n+1}+\ldots+w_N).$$

Let t_n be the midpoint of the nth interval, i.e.

$$t_n = s+w_1+\ \ldots\ +w_{n-1}+(w_n/2),$$

and let RT_n denote the RT when Σ arrives in the nth interval. By the RT-foreperiod curve will be meant the plot of RT_n against t_n.

Almost any shape of this curve can be explained by making appropriate assumptions about the finite sequence $\{w_n\}$. However, consideration will be restricted to sequences of the type $w_n=\alpha w_{n-1}$. Then the following theorem can be proved.

THEOREM 4.4

If $w_n=\alpha w_{n-1}$, then the RT-foreperiod curve is decreasing and is convex if and only if $\alpha \geqslant 1$.

PROOF

(i) If $\alpha=1$, it is easily seen that this case is similar to the discrete case considered at the beginning of Section 4.2 where it was shown that

$$RT_n = a_1 - b_1 t_n.$$

(ii) $\alpha<1$; $w_n=\alpha^{n-1}w_1$.

$$\therefore\quad h=w_1(1-\alpha^N)(1-\alpha)^{-1},$$

and

$$w_n+\ldots+w_N = w_1(\alpha^{n-1}-\alpha^N)(1-\alpha)^{-1}.$$

$$\therefore\quad p_n=(1-\alpha)(1-\alpha^{N-n+1})^{-1},$$

and

$$RT_n = a_2+b_2(1-\alpha^{N-n+1}).$$

Let $\Delta_n f$ denote $f_{n+1}-f_n$. Then,

$$\Delta_{n-1}t = (w_{n-1}+w_n)/2 = \alpha^{n-2}w_1(1+\alpha)/2,$$

and

$$\delta_n = \frac{\Delta_{n-1}RT}{\Delta_{n-1}t} = -2b_2\alpha^{N-2n+3}(1-\alpha)[w_1(1+\alpha)]^{-1} < 0,$$

$$\Delta_n\delta = -2b_2\alpha^{N-2n+1}(1-\alpha)^2/w_1 < 0.$$

That is the RT-foreperiod curve is concave, decreasing.

(iii) $\alpha > 1$; $p_n = (\alpha-1)(\alpha^{N-n+1}-1)^{-1}$;

and

$$\Delta_{n-1}t = \alpha^{n-2}w_1(1+\alpha)/2.$$

Here it can be seen that

$$\delta_n = -2b_2\alpha^{N-2n+3}(\alpha-1)[w_1(1+\alpha)]^{-1} < 0,$$

and

$$\Delta_n\delta = 2b_2\alpha^{N-2n+1}(\alpha-1)^2/w_1 > 0,$$

i.e. the RT-foreperiod curve is convex, decreasing. This proves the theorem.

The convex shapes obtained by Drazin (1961) suggest that a value of $\alpha \geqslant 1$ would be appropriate in this type of model. This would mean that if S adjusts his SR intermittently as he waits for Σ, then his rate of adjustments decreases as he waits.

5. An Alternative Treatment Using the Expectancy Function

In this section it will be assumed that the foreperiod distribution is continuous; it will be seen that analogous results hold for the discrete case. Let $f^*(t)dt$ be the probability that Σ arrives in the infinitesimal interval $(t, t+dt)$. Then, as in Section 3, the expectancy function is defined by the relation

$$p^*(t) = \frac{f^*(t)}{1-F^*(t)}, \quad \text{where } F^*(t) = \int_{-\infty}^{t} f^*(u)du.$$

This mathematical expression is known as the *hazard* function in reliability theory, as the *intensity* function in extreme value theory and as the *force of mortality* in actuarial work. It has already been introduced in reaction-time studies by Audley (1963) as a measure of the theoretical expectancy of an event at any point in the distribution.

In Section 4 we assumed that S's error was in time-keeping, but Audley (personal communication) has pointed out that subjective probability distributions can be significantly different from the corresponding objective ones and some consideration should be given to this fact. Let $f(t)$ be the subjective estimate of $f^*(t)$, and let $p(t)$ be the corresponding subjective expectancy function. In accordance with the arguments outlined in Section 3 it will be assumed that the RT to a signal presented in $(t, t+dt)$ is an increasing function of $1/p(t)$. Then the shapes of the empirical RT curves suggest that, in general, $p(t)$ is an increasing function of t. Two exceptions are the case when $p'(t) \equiv dp(t)/dt$ is negative for small t and then becomes positive (Elithorn and Lawrence, 1955; Klemmer, 1956), and that when $p'(t)$ is positive up to some point and is negative towards the end of the (finite) foreperiod range (Drazin, 1961). The question that will concern us here is under what mathematical conditions on $f(t)$ can these two special cases arise.

Properties of monotone $p(t)$ have been studied by Barlow *et al.* (1963), who showed that if the distribution range is finite, $(0, h)$ say, then

$$\lim_{t \to h} p(t) = \infty,$$

which implies that $p(t)$ cannot be decreasing near h. In other words, a necessary condition for $p(t)$ to be decreasing at the end of the foreperiod range is that the total subjective probability up to the endpoint is less than unity. This is very likely to be the case when a proportion $1-\pi$ of catch trials is included, since then the total objective probability up to the endpoint is $\pi < 1$. It follows that to outline a general approach to the question for all π it is desirable to assume that the subjective foreperiod range is larger than the actual range for all π, including $\pi = 1$. This assumption will be made, although the main interest is in the behaviour of $p(t)$ in $(0, h)$, the interval in which Σ can be presented. It is implied that S always *underestimates* unit probability.

If it is assumed that $RT(t)$ is a linear function of $1/p(t)$, then the RT curves for an ideal S and a uniform foreperiod distribution are straight lines. However, linear RT curves do not necessarily reflect that S is ideal. For, from the definition of $p(t)$,

$$p(t) = -\frac{d}{dt} \ln [1 - F(t)],$$

whence

$$1 - F(t) = \exp [-\int_{\infty}^{t} p(x) dx]. \tag{5.1}$$

Putting

$$p(x) = \frac{\xi}{\eta - x}, \quad \xi > 0, \quad 0 < x < \eta,$$

in eqn. (5.1), it follows that

$$1 - F(t) = \exp [\xi \ln (\eta - t) + \text{constant}].$$

Using the condition $F(0) = 0$,

$$1 - F(t) = \left(1 - \frac{t}{\eta}\right)^{\xi}, \qquad 0 < t < \eta. \tag{5.2}$$

So that if S's estimate of a uniform distribution belongs to the family (5.2), his RT curves would be indistinguishable from those of the ideal S.

The more familiar distributions such as the family (5.2), the normal, the gamma and the Weibull distributions, all yield monotone expectancy functions. It is possible, however, to obtain non-monotone expectancy functions by 'mixing' these distributions. Two such examples which exhibit one or both of the features, i.e., $p(t)$ decreasing for small t, and for t close to h, the endpoint of the range. ‘Mixtures’ of the exponential distribution with (5.2) will be considered for $\xi=1$ and $\xi=2$.

Suppose that

$$\begin{aligned} f(t) &= \frac{\pi_1}{\eta} + \pi_2\lambda\,\mathrm{e}^{-\lambda t} \qquad 0<t<\eta, \\ &= \pi_2\lambda\,\mathrm{e}^{-\lambda t}. \qquad \eta<t. \end{aligned} \tag{5.3}$$

$\pi_1+\pi_2=1$, and it is presumed that $|\eta-h|$ is small.

Then

$$\begin{aligned} p(t) &= \frac{\pi_1+\pi_2\eta\lambda\,\mathrm{e}^{-\lambda t}}{\pi_1(\eta-t)+\pi_2\eta\,\mathrm{e}^{-\lambda t}} \qquad 0<t<\eta \\ &= \lambda \qquad \eta<t. \end{aligned}$$

It can be seen that $p'(0)>0$ if and only if $(\eta\lambda-1)^2<\pi_2^{-1}$, whilst $p'(\eta)$ is always positive, so that this expectancy function can exhibit the first feature (compare Figure 3).

Suppose now that

$$\begin{aligned} f(t) &= \frac{2\pi_1(\eta-t)}{\eta^2} + \pi_2\lambda\,\mathrm{e}^{-\lambda t} \qquad 0<t<\eta \\ &= \pi_2\lambda\,\mathrm{e}^{-\lambda t} \qquad \eta<t. \end{aligned} \tag{5.4}$$

Then

$\eta^4 p'(t)[1-F(t)]^2 = 2\pi_1^2(\eta-t)^2 - \pi_1\pi_2\eta^2\,\mathrm{e}^{-\lambda t}[\{\lambda(\eta-t)-2\}^2-2]$; so that $p'(0)>0$ if and only if $(\eta\lambda-2)^2<2\pi_2^{-1}$, whilst $p'(\eta)$ is always negative.

To show that $p'(t)$ is sometimes positive in $(0, \eta)$ if $(\eta\lambda-2)^2>2\pi_2^{-1}$, it is sufficient to show that $[\lambda(\eta-t)-2]^2-2$ is sometimes negative in this range.

$[\lambda(\eta-t)-2]^2-2=(\lambda t)^2-2(\eta\lambda-2)\lambda t+\eta^2\lambda^2-4\eta\lambda+2$. This expression is negative only between the values of t given by

$$t=\eta-\frac{2\pm\sqrt{2}}{\lambda}$$

and both values lie in $(0, \eta)$ since

$$\eta\lambda-2>\frac{\sqrt{2}}{\pi_2}>\sqrt{2}.$$

This expectancy function can therefore exhibit both features if $(\eta\lambda-1)^2>2\pi_2^{-1}$ and $|\eta-h|$ is sufficiently small.

To study the effect of introducing catch trials we can examine the effect on $p(t)$ of reducing $F(h)$, which can be done by changing any one of the parameters π_1, η or λ. In Figure 4 the effect is shown of reducing π_1 from 3/4 to 1/4 when $\lambda=1/2$, $\eta=4$, $h=3{\cdot}3$; $F(h)$ is reduced from 0·93 to 0·85. It is seen that the effect is to produce a minimum RT at an interval less than the endpoint of the foreperiod range. Similar results would hold for the discrete case if one were to replace the exponential distribution by the geometric and regard t in eqn. (5.2) as a discrete variable.

S.P.

One remaining question concerns the interpretation of a mixed distribution such as (5.3) or (5.4) when the experimenter uses a uniform foreperiod distribution. One can suppose that the continuous treatment given here is an approximation to the following discrete one. Suppose that the time axis is quantized into small intervals of length Δ. In the ith interval S estimates the unconditional probability, $f(i\Delta)\Delta$, that Σ is presented in that interval, and the cumulative sum of his estimates up to the ith interval, i.e. $F(i\Delta) = \sum_{r=1}^{i-1} f(r\Delta)\Delta$. His SR is based on the ratio $\Delta f(i\Delta)/[1-F(i\Delta)]$. One can suppose further that the estimates $f(i\Delta)$ are independent of each other, and result from S's being in one of two states. With probability π_1, S is in the first or 'attentive' state and he behaves like the ideal S, i.e. his subjective distribution is one of (5.2). With probability π_2 he is in the other 'inattentive' state, and in this state his estimate comes from an exponential distribution. With this interpretation the introduction of catch trials should increase the probability of his being in the second state, i.e. reduce π_1, while leaving unchanged the parameters of the component distributions. This is the state of affairs illustrated in Figure 4. On the other hand, one can regard a mixed distribution merely as an approximation to the

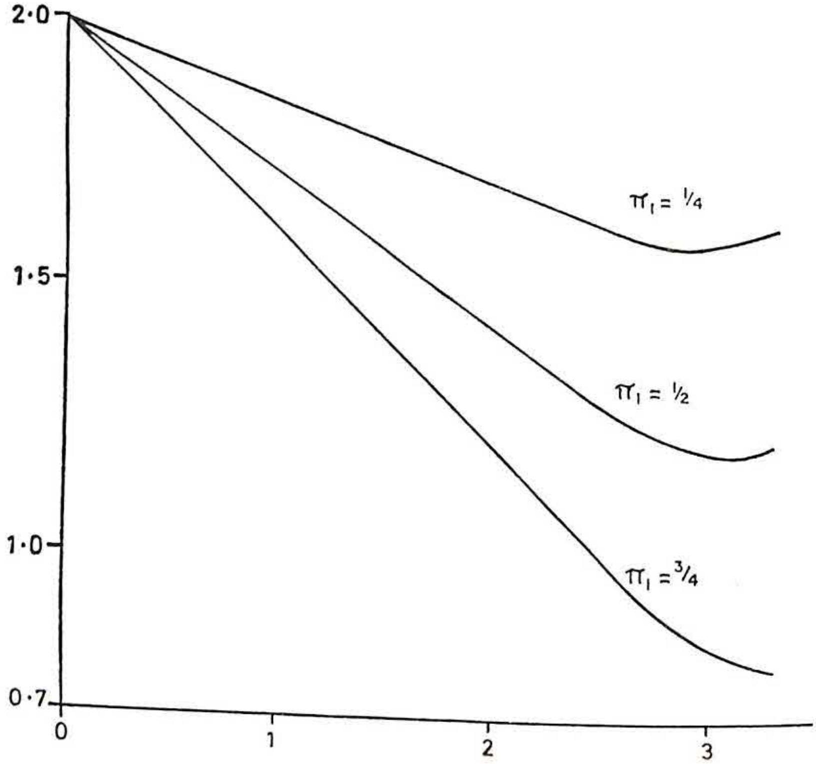


FIGURE 4. Effect of 'catch' trials. Plot of $1/p(t)$ for
$f(t) = 2\pi_1(a-t)/a^2 + \pi_2\lambda\, e^{-\lambda t}$, $0<t<a$,
$= \pi_2\lambda\, e^{-\lambda t}$, $a \leqslant t$.
$\pi_1 = 3/4,\ 1/2,\ 1/4$; $a=4$; $\lambda = 1/2$.

actual subjective distribution. In this case the search should continue for better approximations to the subjective distribution which also give non-monotonic $p(t)$. It may be as well to remark that a ' mixture ' of the two gamma distributions, $\lambda_1^2 t \exp(-\lambda_1 t)$ and $\lambda_2 \exp(-\lambda_2 t)$, can exhibit either of the two features depending on the value of λ_1/λ_2 but because of the absence of data on subjective distributions this question will not be pursued further here.

6. CENTRAL INTERMITTENCY

6.1. *The psychological refractory period*

In the serial reaction task S is told to respond to two signals at each trial. The first, Σ_1, is presented at time t after a warning signal, and the second, Σ_2, after a time $t+I$; I is the inter-signal interval and t and I are variable. The statistic of primary interest here is $RT^{(2)}$, the RT to Σ_2. It has been found that $RT^{(2)}$ is minimal when I is large, that this minimum is maintained as I is reduced down to some value, I_x, and that for $I<I_x$, $RT^{(2)}$ is greater than the minimum (Craik, 1948; Hick, 1948; Vince, 1948). The period of length I_x is called the Psychological Refractory Period. It is the period following the presentation of Σ_1 during which a second signal would give rise to a second RT longer than the minimum. The lengthening of $RT^{(2)}$ above the minimum when $I<I_x$ is called the delay and is generally interpreted by workers in this field as the ' storage ' time, or the interval between the presentation of Σ_2 and the initiation of the second response.

Evidence has been adduced to show that the delay occurs in the central nervous system and the first theory about its magnitude was put forward by Hick (1948) and Welford (1952, 1959). According to Welford, delays are due " to the central processes concerned with two separate stimuli not being able to coexist, so that the data from a stimulus which arrives while the central mechanisms are dealing with data from a previous stimulus have to be ' held in store ' until the mechanisms have been cleared " (Welford, 1952, p. 3). This statement is also referred to as the intermittency hypothesis. So that if $RT^{(1)}$ is the actual first reaction time and RT_m the minimum second reaction time, then, by Welford,

$$RT^{(2)} = RT_m + RT^{(1)} - I.$$

This formula cannot explain delays which occur when $I > RT^{(1)}$. Welford has explained such delays by supposing that before S attends to Σ_2 he has to wait for a feedback signal indicating that R_1 is being made. Davis (1957) has explained similar delays by supposing that there is a ' Central Refractory Period ' after the central processes have dealt with Σ_1 and during which they remain unoccupied.

However, in other experiments (Elithorn and Lawrence, 1955; Halliday *et al.*, 1960; Borger, 1963) the delays obtained were much smaller than either the Welford theory or the Davis refinement would predict. Elithorn and Lawrence attempted to account for the delays by proposing an ' expectancy ' theory in which RT was a function of SR. Their approach, though having its shortcomings (see Davis, Hick and Welford, 1956), will be found to be similar to the one to be adopted here.

The present contention is that temporal uncertainty of signal influences RT; for R_1 the uncertainty is a function of t, and for R_2 it is a function of I. That is, Σ_1 provides the warning signal for Σ_2. In most of the work on intermittency the minimum RT is referred to as the *normal RT*, and any increase over this minimum is called the delay. In the present study the *normal RT* will be taken to be the RT to a signal *when no overt response is required to the warning signal.* Thus, in the present context, the normal second reaction time, $RT_x^{(2)}$ (or simply RT_x), is the RT to Σ_2 when S is not required to respond to Σ_1. To avoid confusion, the term *lag* will be introduced to refer to the increase in RT over RT_x when S has to respond to Σ_1. The lag in the second RT is $(RT^{(2)} - RT_x)$ and measures the response interaction when both signals require a response. It is zero, by definition, either when no response to Σ_1 is required or when I, the inter-signal interval, is greater than $RT^{(1)}$. Positive lags have been obtained by Davis (1959) and Kay and Weiss (1961), who performed experiments in which no response to Σ_1 was required, and in which S responded to both signals.

In sum, it is supposed that S attends to a response as soon as the appropriate signal is presented and that his RT depends on

(*a*) his SR at the time of presentation of Σ, and

(*b*) whether or not he is making another response at the same time.

Some of the theory developed in Section 4 will now be applied to calculate the lags from the data of Davis (1956) and Halliday *et al.* (1960).

6.2. *Calculations using simple and disjunctive RTs*

Before lags can be estimated from observed $RT^{(2)}$ in the serial reaction task, it is necessary to obtain estimates of the normal RT under the given experimental conditions. Let us suppose that Σ_2 can be presented at one of the time-points $t=1, 2, \ldots, n$ (the *clock*-time between the time of presentation of Σ_1 and one of these time-points is simply the inter-stimulus interval, I), and let RT_t denote the normal RT to Σ_2 when it is presented at the tth point. We assume that RT_t is given by

$$\begin{aligned} RT_t &= a + bp_t^{-1} \qquad & 0 < p_t \leqslant p_x \\ &= RT_m & p_t \geqslant p_x, \end{aligned} \tag{6.2.1}$$

where p_t is the expectancy function and can be calculated from a knowledge of the experimental procedure. The constants a, b and RT_m can be determined by fitting eqn. (6.2.1) for values of t corresponding to values of $I > RT^{(1)}$, since for these values of I the lag is, by definition, zero and the observed $RT^{(2)}$ is the normal RT to Σ_2. The fitted curve can then be extrapolated to give estimates of RT_t for all t.

Two remarks will be made concerning the extrapolation of RT curves. First, it was stated in Section 3 that there may be an upper limit, p_x, beyond which greater conditional probability of signal occurrence does not affect RT. So that fitted RT curves should be used as prediction formulae for large p_t only when RT_m is known. Since RT_m is probably in the region $0{\cdot}15-0{\cdot}18$ sec, it will be assumed that if $a+b>0{\cdot}18$ sec then extrapolation of fitted RT curves is valid for all p_t. Second, in the experiments of Halliday *et al.*, Σ was equally likely to be one of two signals. It seems probable that signal detection takes longer in a disjunctive situation than in a simple one, so it will be supposed that the effect of having two or more equally likely signals is to increase a and RT_m in eqn. (6.2.1), while leaving b unchanged. It would follow that if, for a given S, a and b are estimated from disjunctive RTs and then used to estimate normal RT in a simple situation, the values of normal RT so obtained would probably be overestimates.

In the experiments of Davis (1956) a proportion 12/64 of catch trials was included, so that p_t can be calculated by using the theory of Section 4.3 together with knowledge of the experimental procedure. I (secs) was chosen from a 16-point uniform frequency distribution, the points lying between 0·05 and 1 sec. For both Ss examined, RT_t had a minimum at about $t=12$, so using eqn. (4.3.4) and putting $\pi=52/64$, $n=16$ and $t=12$ in eqn. (4.3.2), the estimate of d is 0·028. Also, the estimated value of t (<12) for which RT_t should be approximately equal to RT_{16} is between 5 and 6. This last 'prediction' is borne out by the curves for both Ss. The constants in (6.2.1) were estimated using the method of least squares and the lags, $(RT^{(2)}-RT_t)$, are shown in Table 1.

TABLE 1. LAGS IN $RT^{(2)}$ (IN MILLISECS) FOR DIFFERENT VALUES OF I. $RT^{(1)}$ ABOUT 0·23 SEC. DATA OBTAINED FROM DAVIS (1956)

I (secs)	0·05	0·10	0·15	0·20
Subjects				
S_1	117	69	37	23
S_2	110	25	15	7

In the experiments of Halliday *et al.*, Σ_1 could have appeared on either the left- or right-hand side of a screen and Σ_2 appeared on the side opposite to Σ_1. At each trial signals could have been presented at times 0·00, 0·05 or 0·10 sec $(t=1, 2, 3)$ after the end of a warning signal. Both were presented with probability 3/4, otherwise only one was presented with the restriction that signals were not presented simultaneously. The experimental probabilities of signal occurrence at the three time-points are shown in Table 2 (a). Though catch trials were included, the possible change in subjective belief was likely to be negligible since there were only three possible time-points. Using the probabilities in Table 2 (a) the eqn. (6.2.1) was fitted by plotting RT against probability for the disjunctive single situation only.

One of the conclusions of Halliday *et al.* was that for subject II, $RT^{(2)}$ was 'speeded up' if Σ_2 was a left-hand signal. Values of b and $a+b$ are shown in Table 2 (b); also reproduced there for this subject is $RT^{(2)}$, the *observed* RT

TION'L AND PSYCHOLO

Bureau of Ed'l & Psyl. Research
(S. C. E R T)
Date..

TABLE 2. DATA OBTAINED FROM HALLIDAY ET AL. (1960)

(a) For any 'time' column the element in the leading diagonal is the conditional probability for Σ_1, and those off the diagonal are the probabilities for Σ_2

Times	'0'	'50'	'100'
	7/24	3/7	3/4
		2/5	3/4
			1/2

(b) Estimating a and b by fitting the curve $RT = a + b/p$.

		Right hand		Left hand		
	Period	b	$a+b$	b	$a+b$	$RT^{(2)}$
	1	1·6	239	23·9	227	
S_I	2	19·2	229	17·3	228	
	3	13·1	225	13·1	230	
		Mean	231	Mean	228	
	1	(0)*	217	26·3	192	202
	2	0	233	29·3	214	205
	3	(0)*	219	21·4	212	212
S_{II}	4	2·9	230	15·0	222	237
	5	6·5	218	16·8	225	227
	6	(0)*	233	3·9	235	230
	7	6·7	223	12·7	214	211
		Mean	222	Mean	216	

* Signifies that the estimated value of b is negative.

to the left-hand signal when it came second after an interval of 0·10 sec. A comparison of these values with those of $a+b$ suggests that the small values of $RT^{(2)}$ can be explained in terms of subject II's expectancy. It would still be an open question as to why there seems to be no interaction of responses when Σ_2 is a left-hand signal. The lags are small and are not shown.

Two seemingly interesting practice effects can be seen in Table 2 (b). It appears that in both *S*s $a+b$ is not appreciably affected by practice and that it is about the same for both hands. However, in the unpractised (left) hand of both *S*s there is an approximately linear reduction in b with practice. b is the index of the expectancy effect, and the fact that it is different for both hands suggests that *SR* may be related to muscular tension. One might conclude from these two findings that there is an upper limit to the rate of responding which is resistant to training, but *S* improves his average performance with practice by learning to react more quickly to unexpected stimuli.

7. THE INTERACTION OF TWO RESPONSES

7.1. *Statement and discussion of a model*

The positive lags referred to above will be interpreted as evidence that there is some form of response interaction when the making of two responses overlaps in time. The original statement of the intermittency hypothesis implies the extreme assumption that *S* attends to one response until it is made and then attends to the other. This assumption as it stands is incompatible with the expectancy approach outlined above because the values of the lags obtained are significantly less than $(RT^{(1)}-I)$. However, intermittent behaviour of some sort seems to be a simple explanation of response interaction.

If Σ_2 is presented before R_1 is completed, it can be supposed that S may be able to attend to R_2 for some of the time up to the completion of R_1. This could be the case, for instance, if a response consists of a certain number of stages, some of which are also stages of the other response. Thus when there are two responses to be made, there would be three types of stages, one type comprising stages common to both responses, and the other two comprising those stages belonging to one response and not the other; with these stages being performed serially. Given these three types of stages, response interaction could then be described in terms of the process determining the order in which these stages are performed. This process should take account of at least two features. The first is the possibility that 'preference' is given to the preparation of the stages of R_1, because this response was the first one started. The second is the possibility that a stage of a given response is likely to be followed by a stage of the same response because of some form of 'association' among stages of the same response. The relevant property of a stage is its type, so it is necessary to consider only the 3-state process giving the sequence of types. One such process is the 3-state Markov chain in which states 0, 1, 2 correspond to the types of stages belonging to both R_1 and R_2, to R_1 and not R_2, to R_2 and not R_1, respectively. If the transition matrix is denoted by (q_{ij}), q_{ij} being the probability of going from type i to type j, $i, j=0, 1, 2$, then 'preference' and 'association' would be reflected by the inequalities

$$q_{01} \geqslant q_{02}$$

$$q_{10}+q_{11} \geqslant q_{12}$$

$$q_{20}+q_{22} \geqslant q_{21}.$$

In what follows, the simpler case will be considered where $q_{0j}=q_{1j}=q_{2j}$, for $j=0, 1, 2$.

The organization of a response will be regarded as having a number, n_0 (thought of as being large), of stages which are performed serially, each requiring a random time δ. Because of preparation, when Σ_1 is presented at time t, $n_1'(<n_0)$ stages of R_1 remain to be performed and their execution starts at t. n_1' is a function of S's SR and, therefore, of the temporal uncertainty of Σ_1. Σ_2 is presented at $t+I$ and it is supposed that by that time n_1 $(<n_1')$ stages of R_1 remain to be performed, and that n_2 $(<n_0)$ stages of R_2 remain (because of preparation). The random process which starts at time $t+I$ is assumed to be the following. Starting from the instant that the preparation of a stage is complete, the next stage performed belongs to R_1 (but not R_2) with probability Y_1, to R_2 (but not R_1) with probability Y_2 or to both R_1 and R_2 with probability $(1-Y_1-Y_2)$. The process continues until there are no more stages of a response left, from which time the stages belonging to the other response are performed serially.

The RT data available would allow the moments of δn_i and the values of Y_i, $i=1, 2$ to be calculated. If the Y_i are allowed to vary among Ss, the model can be made to apply to three types of experimental results:

(i) No lags in either $RT^{(1)}$ or $RT^{(2)}$. It could be supposed here that $Y_1=Y_2=0$. In this case the overlapping portions of R_1 and R_2 are performed together or grouped. Vince (1948) has stated that grouping might occur when I is small because S perceives and responds to both signals together. This would lead one to expect simultaneous completion of R_1 and R_2. The present meaning of grouping is somewhat different, and one should not expect R_1 and R_2 to be completed simultaneously since n_1 would probably be less than n_2.

(ii) Lags in $RT^{(2)}$ but not in $RT^{(1)}$. It could now be supposed that $Y_2=0$. If R_1 is completed before R_2 and when R_1 is completed m_1 stages of R_1 (but not R_2) and (n_1-m_1) stages of both R_1 and R_2 have been prepared, then m_1 is a binomial random variable with parameters Y_1 and n_1. Writing $I=E[\delta(n_1'-n_1)]$, it can be seen that

$$RT^{(2)}=RT_N^{(2)}+Y_1(RT^{(1)}-I),$$

where $RT_N^{(2)}$ is the normal RT to Σ_2 and $RT^{(1)}$ is the observed RT to Σ_1.

(iii) Lags in both $RT^{(1)}$ and $RT^{(2)}$. Here it would be supposed that Y_1 and $Y_2>0$ and $Y_1+Y_2<1$.

In the last two cases intermittency couched in probabilistic terms has been introduced.

7.2. *Approximate distribution of RT*

Suppose that δ has the density function

$$f(x)=\beta\,\mathrm{e}^{-\beta x}.$$

Let L_1 and L_2 denote the lags in $RT^{(1)}$ and $RT^{(2)}$ respectively. Then

$$\begin{aligned}\frac{n_1}{\beta}&=E(RT_N^{(1)})-I, \text{ and } \operatorname{var}(RT_N^{(1)})=\sigma^2_{1N}=\frac{n_1'}{\beta^2}\geqslant\frac{n_1}{\beta^2}\\ \frac{n_2}{\beta}&=E(RT_N^{(2)}), \quad \text{and} \quad \operatorname{var}(RT_N^{(2)})=\sigma^2_{2N}=\frac{n_2}{\beta^2}.\end{aligned} \tag{7.2.1}$$

Let A denote the event that a stage of R_1 (but not R_2) is performed so that $p(A)=1-Y_2$. If the probability that R_1 is completed before R_2 is nearly 1, the number, v' of stages performed up to the completion of R_1 has approximately the same distribution as the waiting time, v, up to the n_1th occurrence of A when n_2 is infinite. The probability generating function (p.g.f.) for v is

$$P(z)=\left[\frac{(1-Y_2)z}{1-Y_2 z}\right]^{n_1}.$$

$\therefore$ the approximate characteristic function (c.f.) for RT_1-I is

$$C(\theta)=P[\phi(\theta)],$$

where $\phi(\theta)=\beta(\beta+\theta)^{-1}$.

$$\therefore \quad C(\theta)=\left[\frac{\beta(1-Y_2)}{\beta(1-Y_2)+\theta}\right]^{n_1}, \tag{7.2.2}$$

which, not surprisingly, is the c.f. for the gamma distribution. From (7.2.1) and (7.2.2) it follows that

$$E(L_1)\doteq\frac{Y_2}{1-Y_2}\,[E(RT_N^{(1)})-I]. \tag{7.2.3}$$

Now let m_1 be the number of stages belonging to R_1 (but not R_2) performed by the time R_1 is completed. If n_2 is infinite, m_1 has the binomial p.g.f.

$$Q(z)=[(1-Y_1)+Y_1z]^{n_1}.$$

It is clear that L_2 is the sum of m_1 negative exponential random variables. Therefore, the approximate c.f. for L_2 is $Q[\phi(\theta)]$, from which is obtained

$$E(L_2)\doteq Y_1[E(RT_N^{(1)})-I],$$

and

$$\operatorname{var}(L_2)\doteq\frac{n_1Y_1}{\beta^2}(2-Y_2). \tag{7.2.4}$$

Since $\operatorname{var}(RT^{(2)})=\sigma^2_{2N}+\operatorname{var}(L_2)$, then, from eqns. (7.2.1) and (7.2.4), the following inequality is obtained:

$$\sigma^2_{2N}\leqslant(RT^{(2)})\leqslant\sigma^2_{2N}+Y_1(2-Y_1)\sigma^2_{1N}. \tag{7.2.5}$$

To find the covariance between RT_1 and RT_2, first define $RT^{(1)\prime}=RT^{(1)}-I$. Given that $v=j$,

$RT^{(2)}=RT^{(1)\prime}+$the sum of $n_2-(j-m_1)$ negative exponential random variables which are independent of $RT^{(1)\prime}$.

$$\therefore \quad E(RT^{(1)\prime}RT^{(2)}|v=j)=E(RT^{(1)\prime 2}|v=j)+\beta^{-1}E(RT^{(1)\prime}|v=j)E(n_2-j+m_1)$$

$$=\frac{j}{\beta^2}(1+n_2+Y_1n_1).$$

$$\therefore \quad E(RT^{(1)\prime}RT^{(2)})=\frac{E(RT^{(1)\prime})}{\beta}(1+n_2+Y_1n_1).$$

Noting that $E(RT^{(2)})=\beta^{-1}[n_2+E(m_1)]$

$$=\beta^{-1}(n_2+Y_1n_1),$$

and approximating β^{-1} by $\sigma^2_{2N}/E(RT_N^{(2)})$, it is found that

$$\operatorname{cov}(RT^{(1)\prime},RT^{(2)})=\operatorname{cov}(RT^{(1)},RT^{(2)})$$

$$=\frac{E(RT^{(1)})-I}{E(RT^{(2)})}\cdot\sigma^2_{2N}. \tag{7.2.6}$$

Y_1 and Y_2 can be estimated from the data using eqns. (7.2.3) and (7.2.4), and certain crude checks of the model can be devised. The first is stated in the inequality (7.2.5). The left-hand inequality is satisfied in most of the experiments of Halliday *et al.* and Borger, but the right-hand inequality is not.

Also in the data of Borger, L_2 is sometimes greater than $(RT^{(1)} - I)$ which is inconsistent with (7.2.4). However, because of the large variability usually present in RT data, it is felt that these inconsistencies do not necessarily reflect the inadequacy of the model. This could perhaps be better assessed by comparing empirical and theoretical distribution functions.

8. Concluding Remarks

The first concern of the present study has been to treat mathematically the influence of stimulus probability on RT, and a prerequisite of the approach is the existence of a subjective correlate of temporal uncertainty of the signal. The suggestion has already been made (Audley, 1963) that the appropriate measure of temporal uncertainty is the conditional probability that a signal would be presented given that it has not yet been presented. This conditional probability has been called here the expectancy function and it has been the variable most exploited in the analysis. The experiments of Walter *et al.* (1964) and Sutton *et al.* (1965) suggest that a subjective correlate of signal uncertainty, SR, can be found in the cortical activity of the subject. It is clear, however, that there is a need for more empirical work designed to determine both the nature of SR and its relation to the expectancy function, on the one hand, and RT on the other.

In Sections 4 and 5 ways have been examined of incorporating subjective aspects such as time and probability estimation into a theoretical study of RT curves. In the approach, use has been made of a distinction between those situations in which the foreperiod distribution is discrete and those in which the distribution is continuous. In the former situations the points of the distribution are points at which SR reaches local maxima, and it follows from this that errors in time estimation lead to reduced preparedness. On the other hand, lower preparedness cannot be the result of errors in probability estimation alone unless the subjective probability estimate has a negative bias. If, however, the foreperiod distribution is continuous with finite range, there are no preferred time-points at which SR should be maximized and within the foreperiod range the idea of time estimation seems irrelevant. To be sure, the beginning of the range has to be estimated, but after then the subjective transformation, $f(t)$, of the experimental probability density function, $f^*(t)$, is essentially a probability estimate at t, and it is this transformation which, by hypothesis, determines RT. It is not argued that discrete and continuous distributions are necessarily distinct psychologically, but only that if they are, then this distinction is important to the theorist, because his assumptions about SR are different in the two cases, as is the relative significance of time and probability estimation.

Two ways of treating the case of continuous foreperiod distributions have been discussed. One approach (Section 4.4) has been to assume that S adjusts his SR intermittently, as in the discrete case, and that his SR remains constant between the times of adjustment. It is shown that the existing empirical RT curves would, on the basis of this model, imply that S's rate of adjusting SR

decreases with time. The other approach, as outlined in Section 5, is analogous to the approach used for the discrete case. It assumes that $RT(t)$ is a linear function of $p(t)^{-1}$, where $p(t)$ is the continuous expectancy function.

The other topic for which two ‘ competing ’ treatments have been offered is that of catch trials. In section 4.3 time-estimation errors were neglected and it was assumed that probability estimation is affected by changing subjective belief in the hypothesis that a trial is a catch trial. In the other treatment, mentioned in Section 5, it was assumed that probability estimation is the result of a ‘ mixing ’ of the objective foreperiod distribution and another distribution, and that the introduction of catch trials changes certain parameters of this ‘ mixed ’ distribution. The assumptions made about subjective belief and about ‘ mixed ’ distributions have, at present, no empirical justification, but it is felt that the theoretical analysis may have some value if only because it demonstrates the possibility of accounting for certain experimental facts.

Studies of reactions to discrete pairs of signals arose out of attempts to analyse human limitations in continuous performance tasks. Because of this, it is perhaps not surprising that the intermittency hypothesis is a statement about the limitations of the central nervous system. Thus the first signal, Σ_1, is regarded as highly ‘ compelling ’ from *S*'s point of view, having a distracting effect and requiring some sort of discrimination whether or not a response to it is required. The central mechanisms would be ‘ captured ’ by Σ_1, and this would give rise to a delay in $RT^{(2)}$ since no two central organizing times can overlap (Davis, 1959). In the expectancy approach adopted here, Σ_1 is ‘ compelling ’ only in so far as it provides the last warning signal after which *S* can expect Σ_2, and it is not assumed that the central mechanisms are necessarily limited to dealing with one signal at a time. So the intermittency studies differ from the *RT*-foreperiod studies only because a response to Σ_1 is required, and to see if this difference affects $RT^{(2)}$ the difference between the *RT* to Σ_2 in the two situations can be measured. This difference, which has been called the lag, then measures the degree of interaction between the making of two responses. The model for response interaction which has been proposed is based on the idea that the central mechanisms while attending to one response can, for some of the time, attend to another response as well. This idea is in a sense a probabilistic analogue of the intermittency hypothesis, and by varying the probabilities involved, the model can account for different types of reactions. This is another example in which a theoretical model is useful in accounting for qualitative features of existing *RT* data, and this has been the main aim of this paper.

ACKNOWLEDGEMENTS

The author wishes to thank Miss V. R. Cane for helpful discussion, and Dr. A. T. Welford for his encouragement and comments on central intermittency. The present form of the paper is due largely to the valuable suggestions of the referees and to a stimulating discussion with Professor R. J. Audley. Acknowledgement is also given of the financial assistance received from the Government of Guyana and the University of the West Indies.

REFERENCES

AUDLEY, R. J. (1963). Decision models in reaction time. In: Psychophysics and the ideal observer. *Proc. 17th Int. Congr, Psychol* pp. 236–237. Amsterdam : North Holland.

BARLOW, R. E., MARSHALL, A. W. and PROSCHAN, F. (1963). Properties of probability distributions with monotone hazard rate. *Ann. math. statist.* **34,** 375–389.

BERTELSON, P. and BOONS, J. P. (1960). Time uncertainty and choice reaction times. *Nature,* **187,** 531–532.

BERTELSON, P. (1966). Central intermittency twenty years later. *Q. J. exp. Psychol.* **18,** 153–163.

BORGER, R. (1963). The refractory period and serial choice reactions. *Q. J. exp. Psychol.* **15,** 1–12.

CRAIK, K. J. W. (1948). Theory of the human operator in control systems. II. Man as an element in a control system. *Br. J. Psychol.* **38,** 142–148.

DAVIS, R., HICK, W. E. and WELFORD, A. T. (1956). Comments on ' Central Inhibition: Some Refractory Observations ' by A. Elithorn and C. Lawrence. *Q. J. exp. Psychol.* **8,** 39–41.

DAVIS, R. (1956). The limits of the ' psychological refractory period '. *Q. J. exp. Psychol.* **8,** 24–38.

DAVIS, R. (1957). The human operator as a single channel information system. *Q. J. exp. Psychol.* **9,** 119–129.

DAVIS, R. (1959). The role of ' attention ' in the psychological refractory period. *Q. J. exp. Psychol.* **11,** 211–220

DAVIS, R. (1965). Expectancy and Intermittency. *Q. J. exp. Psychol.* **17,** 75–78.

DRAZIN, D. H. (1961). Effects of foreperiod, foreperiod variability, and probability of stimulus occurrence on simple reaction time. *J. exp. Psychol.* **62,** 43–50.

ELITHORN, A. and LAWRENCE, C. (1955). Central inhibition: some refractory observations. *Q. J. exp. Psychol.* **7,** 116–127.

HALLIDAY, A. M., KERR, M. and ELITHORN, A. (1960). Grouping of stimuli and apparent exceptions to the psychological refractory period. *Q. J. exp. Psychol.* **12,** 72–89.

HICK, W. E. (1948). Discontinuous functioning of the human operator in pursuit tasks. *Q. J. exp. Psychol.* **1,** 36–57.

HICK, W. E. (1952). On the rate of gain of information. *Q. J. exp. Psychol.* **4,** 11–26.

KARLIN, L. (1959). Reaction time as a function of foreperiod duration and variability. *J. exp. Psychol.* **58,** 185–191.

KARLIN, L. (1965). Comments on ' Expectancy and intermittency '. *Q. J. exp. Psychol.* **17,** 351.

KAY, H. and WEISS, A. D. (1961). Relationship between simple and serial reaction time. *Nature,* **191,** 790–791.

KLEMMER, E. T. (1956). Time uncertainty in simple reaction time. *J. exp. Psychol.* **51,** 179–184.

KLEMMER, E. T. (1957). Simple reaction time as a function of time uncertainty. *J. exp. Psychol.* **54,** 195–200.

MOWBRAY, G. H. (1960). Choice reaction times for skilled responses. *Q. J. exp. Psychol.* **12,** 193–202.

MOWRER, O. H. (1940). Preparatory set—some methods of measurement. *Psychol. Rev. Monog.* No. 52.

SUTTON, S., BRAREN, M., ZUBIN, J. and JOHN, E. R. (1965). Evoked-potential correlates of stimulus uncertainty. *Science*, **150,** 1187–1188.

VINCE, M. A. (1948). The intermittency of control movements and the psychological refractory period. *Br. J. Psychol* **38,** 149–157.

WALTER, W. G., COOPER, R., ALDRIDGE, V. J. and MCCALLUM, W. C. (1964). Contingent Negative Variation: An electric sign of sensori-motor association and expectancy in the human brain. *Nature*, **203,** 380–384.

WELFORD, A. T. (1952). The 'psychological refractory period' and the timing of high speed performance. *Br. J. Psychol.* **43,** 2–19.

WELFORD, A. T. (1959). Evidence of a single-channel decision mechanism limiting performance in a serial reaction task. *Q. J. exp. Psychol.* **11,** 193–210.

WOODROW, H. (1930). The reproduction of temporal intervals. *J. exp. Psychol.* **13,** 473–499.

Vol. 20
Part 1
31–50

The British Journal of Mathematical and Statistical Psychology

May
1967

A DISCRETE PREDICTING CONTROLLER

By Lisanne Bainbridge

Department of Industrial Engineering, University of California, Berkeley [1]

This paper describes an attempt to develop a control device which has some of the characteristics of human processing rather than of the lead-lag formulations of control theory. The controller's sensory limitations, the way in which it internally represents the behaviour of the system under control, and its control strategy, are described. A digital computer simulation of the controller has been subjected to two performance tests; and the limitations of its behaviour are discussed.

1. Introduction

Mathematical models of the human operator controlling a machine or doing a tracking task have previously been made in the language of control theory, which is convenient when the model is used to represent the operator in an overall study of man–machine relations as all the parts can be described in terms with known implications. A control situation is shown in Figure 1. The task is to bring C, the system output value, to the required value A. The controller makes a corrective change of the system input M according to the difference between C and A. A control theory model of the controller is an equation relating the controller input $(C-A)$ to its output M, and is obtained analytically from records of the controller's behaviour. A useful review of this approach is given by Sheridan (1962).

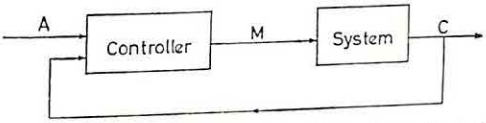


Figure 1. Controller–system relations. Identification of symbols in a car-driving example:

A	target value	required position on road
M	manipulated variable	steering-wheel angle
System		car-steering sub-system
C	controlled variable	actual position on road.

Table 1 compares some of the ways in which automatic and human controllers differ, as shown by experimental evidence. Figure 2 gives an industrial example of human control activity in a papermill (Beishon, 1966, 1967). It is clear that although an automatic control model can behave in much the same way as a human controller the underlying mechanisms are not the same. This paper reports an attempt to synthesize a controller, called FRED, with the human characteristics given in Table 1.

---

[1] Now at the Department of Psychology, University of Bristol.

TABLE 1. COMPARISON OF AUTOMATIC AND HUMAN CONTROLLERS

PROPERTY	*AUTOMATIC CONTROLLER*	*HUMAN CONTROLLER*
	Input Variable Measurement	
controlled variable	any value on continuum	value to nearest display scale unit (e.g. McCormick, 1964)
rate of change of controlled variable	any value on continuum	if rate variable not displayed controller makes absolute judgements limited to 7 ± 2 values (Miller, 1956)
timing of measures	continuous, or sampled at equal time intervals (sampled data controllers)	distribution of sampling rates rather than single rate (Senders, 1964)
	Response Setting	
manipulated variable	any value on continuum	discrete positions which are closer near target value (Table 6)
timing of corrective actions	continuous, or at regular intervals	intermittent (Figure 2)
	Controller Operation	
control type	feed-back, responds to present error	feed-back (Figure 2)
		feed-forward, uses predictions in selecting corrective action (Poulton, 1957)
size of corrective action obtained from	solution of controller equation	?
		chooses values so that controlled variable will not overshoot target (Beishon, 1966 a)

FRED can control any linear slow response over-damped second or higher-order system in response to a single-step change of A. The simulation, which will be described later, contained a second-order system. The response of such a system at steady state (i.e. when the output value has not changed for some time) to a step change in its input M is given by the dotted lines in Figure 3. A higher-order system has the same type of behaviour. If FRED is a useful model it should be made capable of controlling the more complex systems that a human controller can manage.

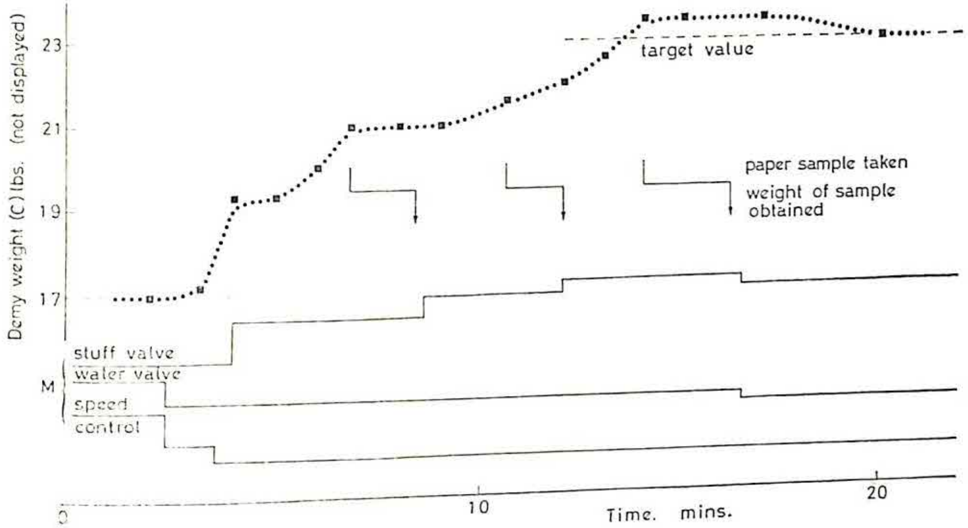


FIGURE 2. Paperman's control actions. Change of paper weight from 17 to 23 lb. From Beishon (1966).

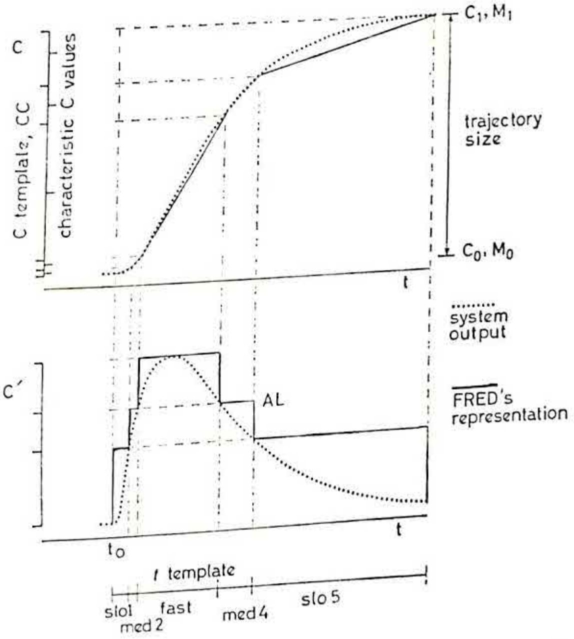


FIGURE 3. Over-damped second-order system behaviour from steady state in response to step change in input from M_0 to M_1 at t_0. FRED's representation of the response is shown; $w = \frac{1}{2}$.

FRED reads the present system output value C_p from the interface, compares this with A and makes an absolute judgement of C', the rate of change of C. It then predicts the effect of various M settings on C and uses these predictions to choose a sequence of M values, and their timing, which will bring C to rest at A in minimum time. Each member of the sequence has a predicted effect on the system which is used as the basis for choosing the next value. As the control sequence allows for predicted behaviour this is feedforward control. When FRED is about to make the next action of the sequence it checks that the system is at the predicted value; if it is, FRED continues, otherwise it chooses a new sequence of M values for control from this C_p. This is feedback control as FRED checks the correctness of its actions.

The basic sequence of M values (R1, R2, R3 and R4) is shown in Figure 4. The top drawing shows the pattern of values chosen when C_p is stationary or moving towards A. R2 moves the system output towards A, R3 brings it to a stop at A, and R4 maintains it at this value. When C_p is moving away from A a first value R1 brings C to a halt at C_T; and from this position R2 is used.

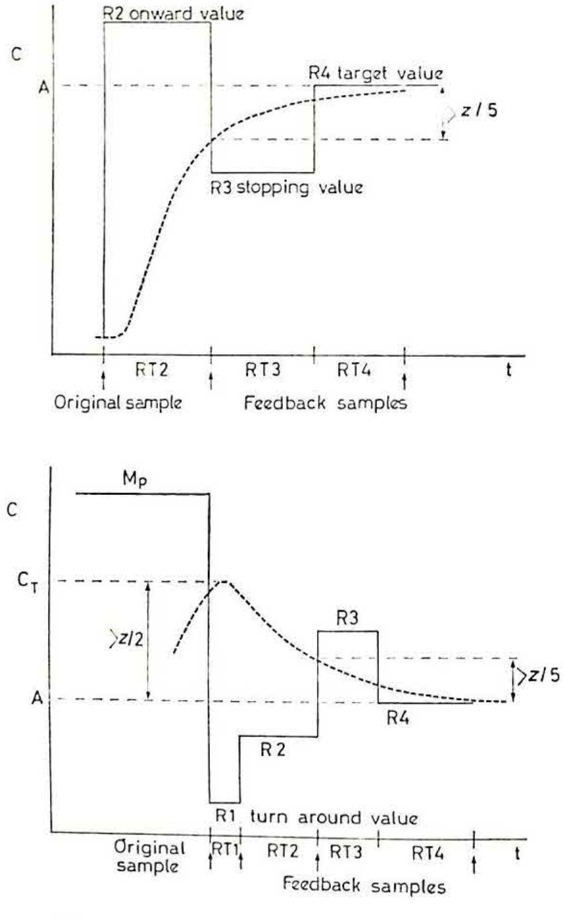


FIGURE 4. Basic controller strategy.

2. CONTROLLER SENSITIVITY FOR INTERNAL VARIABLES

As the system output (C), and time (t), which are continuous variables, are ‘ displayed ’ to FRED it can only obtain their values to the nearest interface unit. C' is not displayed, and FRED obtains its value by comparing C values at successive times. One of FRED's basic features is that it can use only a limited number of *characteristic* values for a variable in its internal operations. The rate value is obtained by internal operations; this is equivalent to the human process of absolute judgement, so the controller is only allowed to distinguish between seven different C' values at any one time—three increasing rates, one stationary and three decreasing rates. Two factors determine the actual values which can be discriminated:

Ratio Rate discrimination in one direction is limited by a Weber fraction, $\Delta C'/C' = w$. A series of physical magnitudes which are discriminably different is given by

$$C_n' = C_0'(1+w)^n, \qquad n = 0, 1, 2.$$

For instance, with $w=1$, actual values which could be distinguished are 1, 2, 4, or 2, 4, 8, or 3, 6, 12, or 4, 8, 16, where the first is called ‘ slow ’ by FRED, the second ‘ medium ’ and the third ‘ fast ’, whatever its actual value. FRED uses $w=\frac{1}{2}$, a value which was chosen arbitrarily. Work is needed on human Weber fractions in this context.

Adaptation level, AL As can be seen from the example, a result of using a ratio basis for discrimination is that the same physical value can fall in a different category in different discriminations. Whether an actual rate of 4 units is called fast, medium or slow depends on the adaptation level of the controller, which FRED selects according to the range of C' values relevant at the time (see Figure 5).

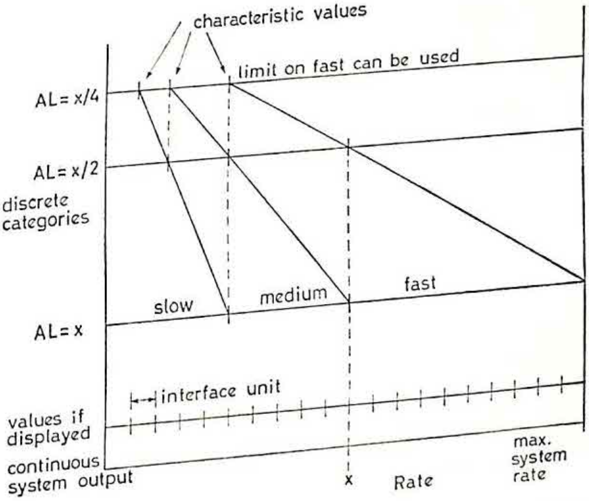


FIGURE 5. Discriminable rate values at three related adaptation levels; $w=1$.

The physical values which can be discriminated at a given AL are the characteristic values used at this time by the controller, and these mark the division of the continuous variable into sections. All actual values in a given section are discriminated as having the characteristic value, which is known by the controller. The AL is arbitrarily specified as the medium rate.

Nesting The effect of the adaptation level is to alter the position of sections on the continuous variable, so changing absolute but not relative accuracy. Figure 5 shows three AL's which are related in such a way that they give a nested succession of discriminable categories. This nesting is made possible by the ratio mechanism. To remain in this succession FRED cannot change adaptation level continuously but only to discrete values that are themselves discriminable. The joint effect of ratio discrimination and AL is that although a limited number of discrete categories are available for discrimination they can be used to cover the entire variable range or to allow fine discrimination, and using discrete adaptation levels maintains the coherence of these changes.

3. Fred's Representation of Future System Behaviour

FRED chooses control actions by predicting the effects of possible actions. To make these predictions it must contain some standard templates of the effect of control actions on system output. As there is no experimental evidence on human templates, FRED's were devised to describe system behaviour given the above limitations on internal accuracy.

SIMPLE TRAJECTORIES

The response of the system at steady state to a step change in input has already been given in Figure 3. For simplicity, at the steady state C value for a given M, $C_0 \equiv M_0$; in real control tasks the transformation from C scale to M scale is one of the sources of task difficulty. When M is changed to another value M_1, C changes until it is at the value corresponding to M_1, $C_1 \equiv M_1$. Whatever the distance between C_0 and C_1, the *size* of the trajectory, C takes the same time to reach its new value, during which the C' changes follow a standard pattern of relative values with time, although the actual values depend on the trajectory size.

As FRED divides the rate variable into discriminably different sections by a ratio we can use this to characterize the standard pattern of C' changes (see Figure 3). For a given size of simple trajectory the rates will range from 0 to some $C'_{\max}$. Taking this maximum as the upper limit of the fast category, the range of rates is sectioned by the ratio; the rate will be slow (slo1) for a certain length of time, then medium (med2), fast, medium (med4), slow (slo5). The length of time spent in each category will be the same whatever the size of the change; this sequence of times t_{1-5} will be called the time (t) template.

The trajectory size, by determining $C'_{\max}$ determines the AL used for discriminating rates during this change. Trajectory size is $C'_{\max}/k$, where k is the system fast-rate constant and $\mathrm{AL} = \sqrt{(k \times \text{size})}$.

A corresponding template for the proportion of the trajectory passed through during each section of the t template can be obtained from the template and rate constants. This is called the C template, CC. The central value of C in each section is used as the characteristic value for C in that section.

$$CC_i = \frac{k}{(1+w)^n} \cdot t_i, \quad (i, n) = (1, 2), (2, 1), (3, 0), (4, 1), (5, 2).$$

and

$$\sum_{i=1}^{5} CC_i = 1.$$

Unfortunately this representation assumes that FRED discriminates C and t perfectly, while it actually knows them only to the nearest interface unit. Part of the slow sections of the time template will be passed while C traverses the first or final unit of a change; as trajectories become smaller med2 and 4 will also be affected, and so on. The t which passes before a change in C can be detected will obviously vary with trajectory size; the above method gives no way of determining this time but does give a way of identifying the trajectory size at which different sections begin to be affected. The *minimum trajectory* z which can be represented correctly, in which C traverses the C display unit (1) during t_1, is given by $1/CC_1$.

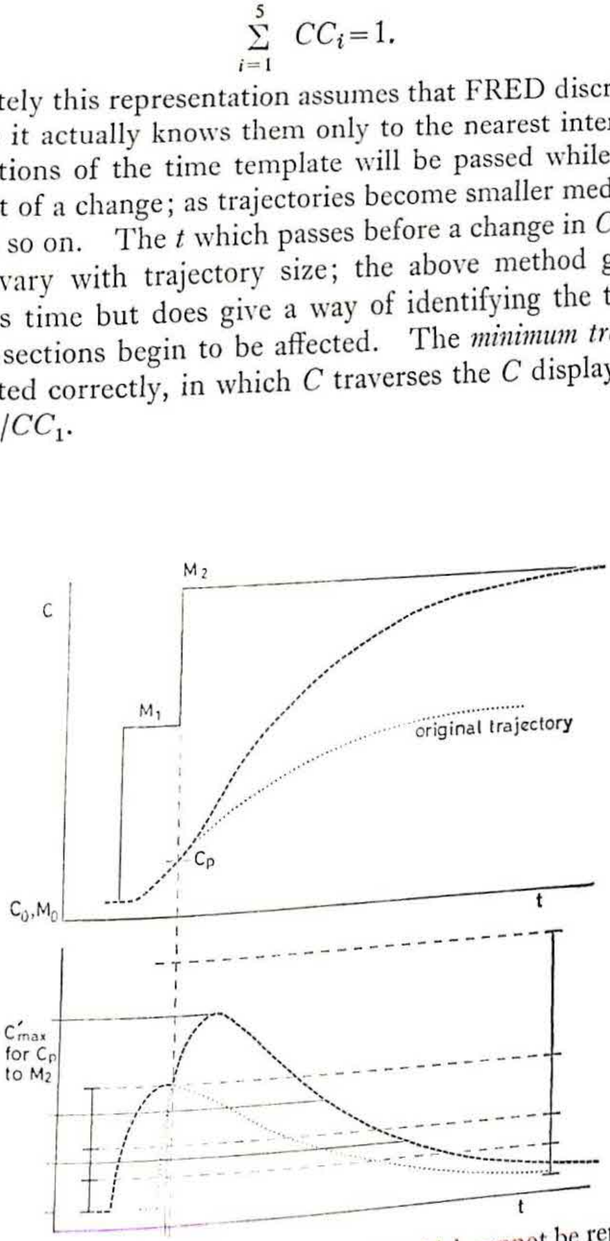


FIGURE 6. An example of an added second trajectory which cannot be represented clearly; $w=1$.

TRAJECTORIES AFTER A STEP CHANGE IN M WHILE C IS CHANGING

There are three situations to be considered:

(1) Addition—the M change to a new value M_2 and movement of C are in the same direction (see Figure 6);

(2) M is moved in the opposite direction to C;

  (*a*) Subtraction – C_p is less than the steady-state value equivalent to M_2 (see Figure 7);

  (*b*) Turn around – C_p is past the steady-state value equivalent to M_2 (see Figure 8).

Addition When M_2 is further from C_p than M_1 the rate pattern followed by C after the change is the same as the rate pattern for a simple control change of size C_p to $C \equiv M_2$, this size determines the maximum rate. The new rate pattern continues from C_p' (see Figure 6).

The two trajectories will be discriminable only if the characteristic value of the rate category of C_p' in the first is the same as that of its rate category in the second, although the categories have different names, as only values in ratio to the characteristic values of the first trajectory can be discriminated. C_p', which may be in any category of the present trajectory, can enter three different next trajectories, as slo1, med2, or fast, so C_p' can be a member of the three adaptation levels

$$C_p'(1+w)^n$$

with $n=1$, 0, -1 respectively. The largest of these trajectories is the one which C_p' joins in the slow section, i.e. $C_p'(1+w)^2/k$. Also, as FRED cannot discriminate between parts of a trajectory falling in one section (see Figure 3) it will

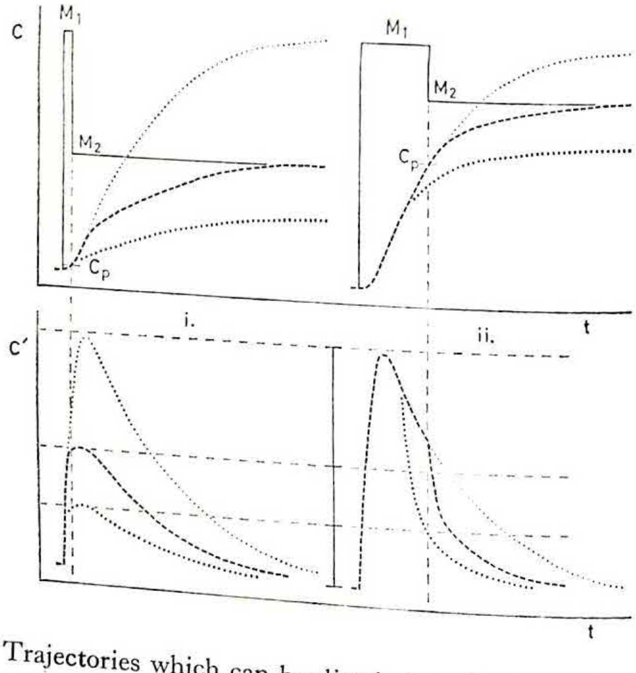


FIGURE 7. Trajectories which can be discriminated when subtracting; $w=1$.

not be able to predict the timing of future behaviour with any accuracy unless the control change is made at a discriminable point in the new trajectory, the junction of two t template sections. While choosing a control sequence FRED is restricted to considering trajectories which it can represent clearly.

Subtraction The effect of reducing the distance between C_p and $C \equiv M$ depends on whether $C'_{\max}$ for the second trajectory has already been passed in the first. If it has not, system behaviour after the change is as for addition (see Figure 7 (i)). If this $C'_{\max}$ has been passed, the new trajectory is joined during deceleration instead of acceleration. The deceleration curve has a standard shape, deceleration from C_p to $C \equiv M$ is the same for a given C_p to $C \equiv M$ distance whatever the original trajectory size. This means that when C_p to $C \equiv M$ is reduced there is a short period of rapid deceleration until the correct C' for the new C_p to $C \equiv M$ is attained and from this value the curve follows the standard form (see Figure 7 (ii)). Arbitrarily, deceleration time is taken as one t unit and the length of C traversed during this period as the average of C_p' and the target rate over the period. Again FRED can only reduce the trajectory by values which will give, at the end of rapid deceleration, a C' value that can be distinguished from C_p', i.e. $C_p'\ (1+w)^n$ for $n=1, 2$, being the medium and slow rates of the trajectory in which $C_p' = C'_{\max}$.

Turn Around If C has passed $C \equiv M_2$, C first comes to a stop, then the trajectory follows the simple form from the turn around point C_T, the C value at which $C' = 0$, to M_2 (see Figure 8).

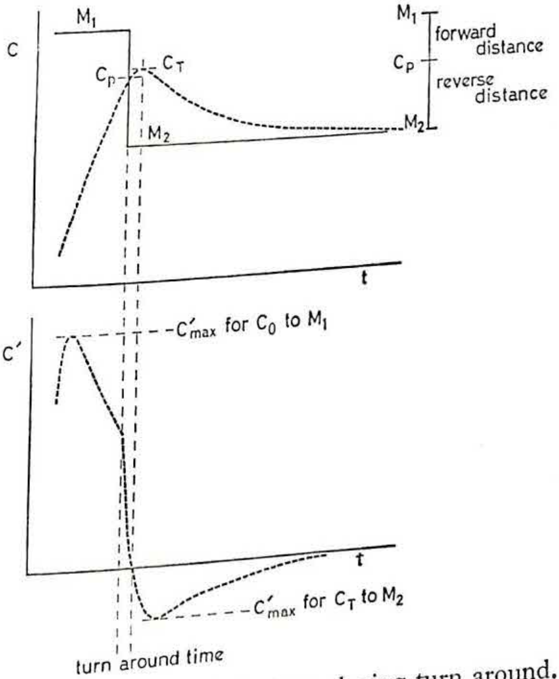


FIGURE 8. System behaviour during turn around.

The turn around time, and the distance from C_p to C_T, are proportional to C_p to $C \equiv M_1/C \equiv M_2$ to C_p. A single value combining both effects is used for each of the t and C turn around parameters, TK and CK. These t and C values are obviously limited in accuracy to the t and C interface units. As this is an internal manipulation it should be even more limited in accuracy, but this will not be investigated here. M_2 values with discriminably different effects could be used in a more elaborate version.

4. THE STRATEGY USED BY THE CONTROLLER

FRED's control strategy has already been described and is shown in Figure 4 and Tables 2 and 3. Some details of this strategy should be mentioned.

Conservatism R1 and R3 are values used to bring C to a stop. They could be end of scale values, but FRED is more conservative and chooses values so that the distance C traverses before $C' = 0$ is greater than a minimum size. Experimental evidence about human behaviour in this situation is needed. The limits were chosen arbitrarily: R1 so that (C_T to A) is greater than $z/2$; R3 so that C is stopped from a distance greater than $z/5$.

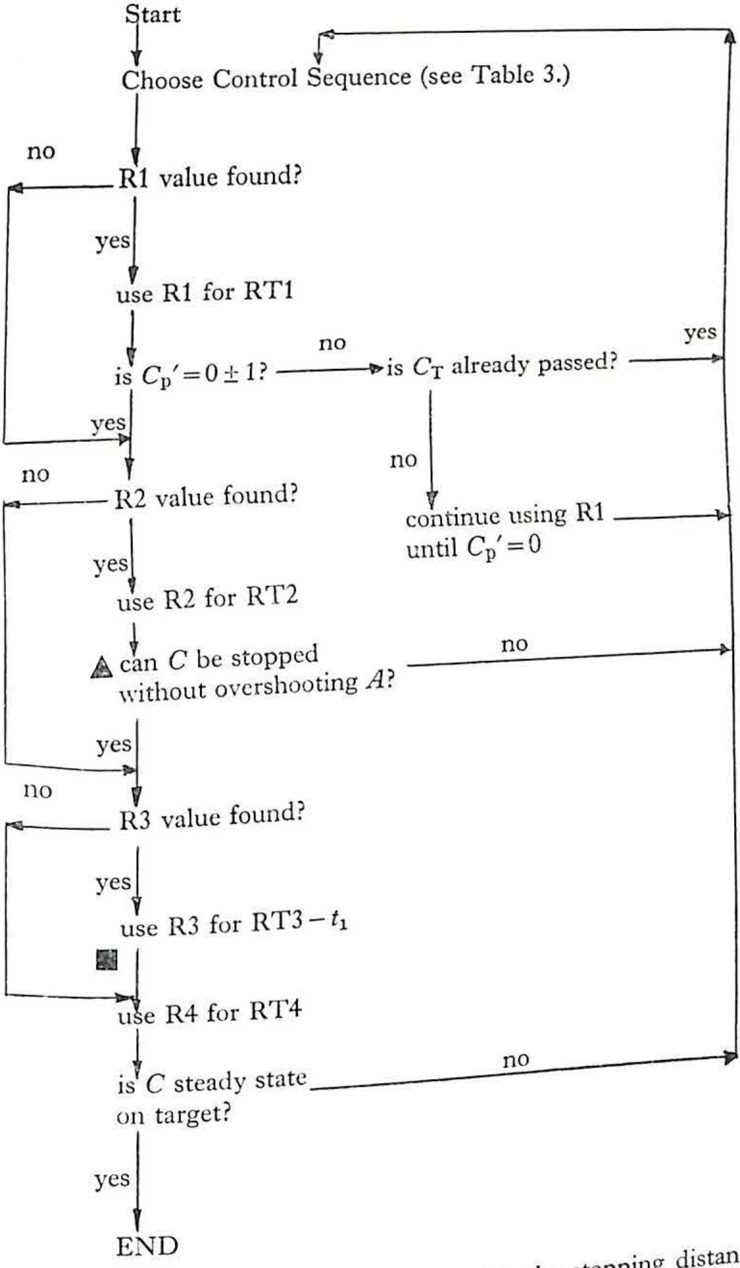
Choice of R2 R2 and R3 are found by predicting the effect of an M (R2) setting in bringing C towards A, and then finding an M (R3) value which will stop C without overshooting A, basing this prediction on the characteristic value of C in the R2 trajectory section which contains A. If no R3 value does this, C values in sections further from A, up to the one containing C_p, are tested in turn until a section associated with an acceptable R3 value is found. R2 is used during control for RT2, the time C takes to pass from its present section in the R2 trajectory to the section in which R3 is imposed, which is calculated using the t template.

There would be many M settings for which an R3 value could be found, but one is required which will bring C to A in the shortest time. FRED therefore first selects a strategy for suggesting M values whose effects will be tested. As there is no evidence about human strategies, FRED's are based on the possibilities allowed by the system representations. These are either to increase C' using the addition mechanism for representing system behaviour or to decrease C' by subtraction, using either adaptation levels in which the present rate can fall, or an adaptation level newly chosen by the controller as relevant to the test, or to use an end of scale M value, which also ignores the present adaptation level and tests the effect of the largest possible change, going to the adaptation level of this trajectory. Deciding which of these strategies is most appropriate depends on the relation between C_p, C_p' and A, as summarized in Table 4.

5. THE CONTROLLER SIMULATION

The controller uses only arithmetic and logical operations, so it is adequately simulated in FORTRAN IV. The system is simulated by expressing its

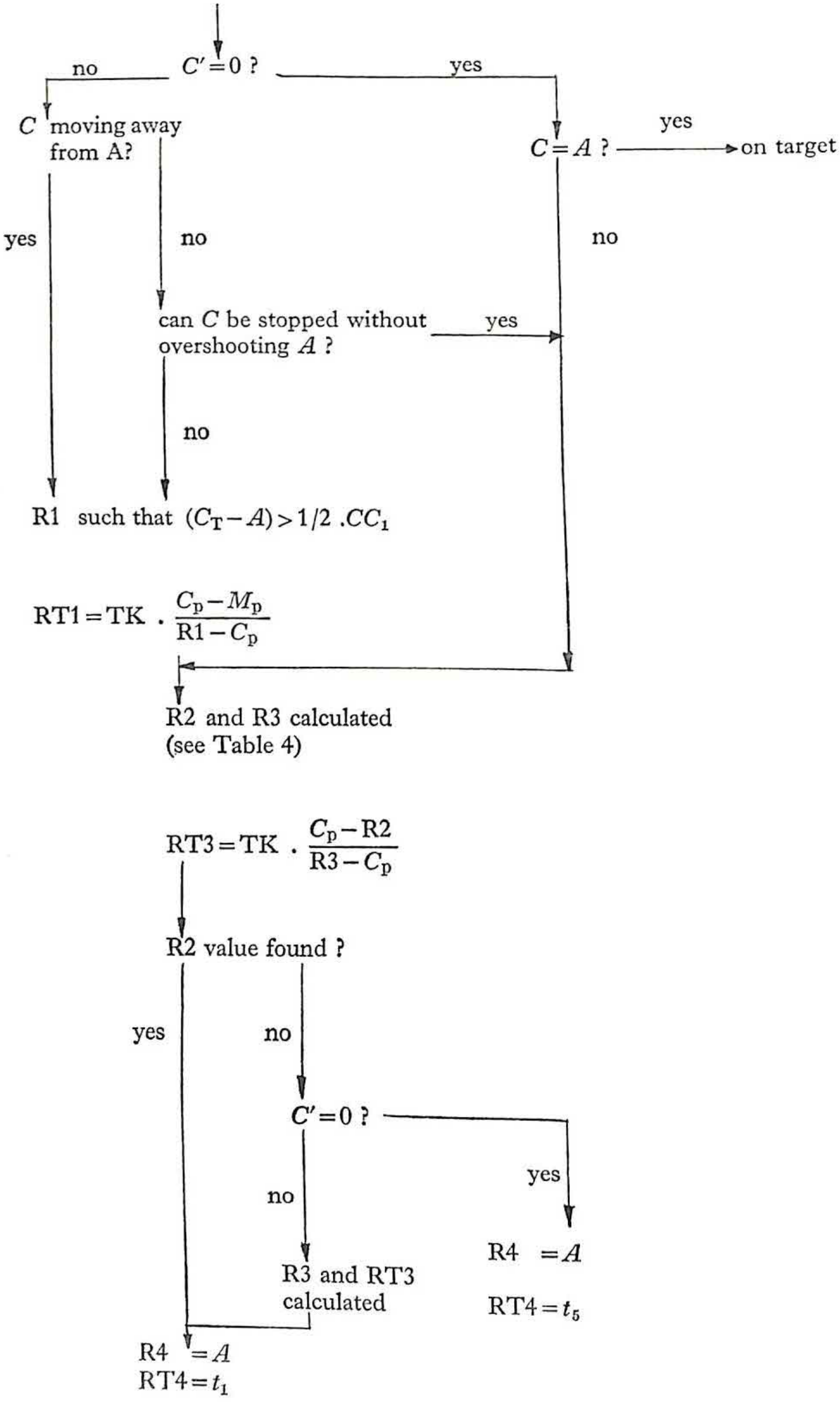
TABLE 2. USE OF CONTROL SEQUENCE



▲ This check is made by comparing (C_p to A) with the stopping distance from C_p specified to a tolerance of the distance traversed in one t unit at the slow rate of this AL. C_p' is categorized as a characteristic value at the R2 trajectory AL.

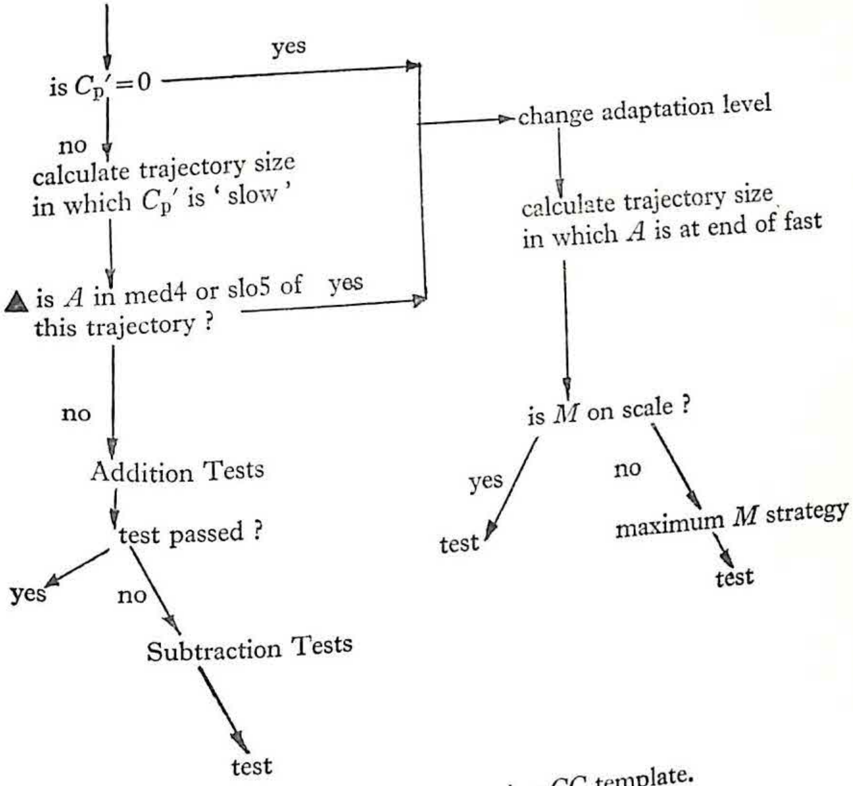
■ Limited C and t discrimination means that some time passes before the effect of an M change on C can be discriminated, so a change should be made at this time before its effect is required. Arbitrarily this time is taken as t_1.

TABLE 3. SELECTION OF CONTROL


$C'=0$?
no
yes
C moving away from A?
$C=A$?
yes
on target
yes
no
no
can C be stopped without overshooting A ?
yes
no
R1 such that $(C_T-A)>1/2 \, . CC_1$
$\mathrm{RT1}=\mathrm{TK} \, . \, \frac{C_p-M_p}{\mathrm{R1}-C_p}$
R2 and R3 calculated (see Table 4)
$\mathrm{RT3}=\mathrm{TK} \, . \, \frac{C_p-\mathrm{R2}}{\mathrm{R3}-C_p}$
R2 value found ?
yes
no
$C'=0$?
no
yes
R3 and RT3 calculated
R4 $=A$
RT4 $=t_5$
R4 $=A$
RT4 $=t_1$

weighting function as a time series and calculating the convolution of this with the M series. The program is in two parts: (I) which calculates the weighting function, and uses this to obtain the system parameters used by FRED; and (II) the simulation of FRED and the system. FRED and the methods for obtaining system parameters in program I are independent of the actual weighting function, and will be effective for any linear slow-response over-damped second-or-higher-order system.

TABLE 4. CHOICE OF M VALUE TO TEST AS R2


is $C_p{}'=0$
yes
no
change adaptation level
calculate trajectory size in which $C_p{}'$ is ‘ slow ’
calculate trajectory size in which A is at end of fast
▲ is A in med4 or slo5 of this trajectory ?
yes
no
is M on scale ?
Addition Tests
yes
no
test
maximum M strategy
test
test passed ?
yes
no
Subtraction Tests
test

▲ Section position found using CC template.

Program I This program calculates the over-damped second-order weighting function $H(t)$ of this simulation as a time series specified at 1 sec. intervals, using the system time constants τ_1 and τ_2. The maximum value of the series is k, with

$$H(t)=\exp(-t/\tau_2)-\exp(-t/\tau_1)$$

and

$$\sum_{t=0}^{\infty} H(t)=1, \qquad (t=1, 2 \ldots).$$

The series is terminated when $H(t)<0{\cdot}01$.

t_{1-5}, measured in seconds, are marked by the positions of the rate constants in the $H(t)$ series (see Figure 3). The t unit is taken arbitrarily as the shortest of the t template sections, measured in seconds. The time template is re-measured in t units for use by the controller, although it must be expressed in seconds for use with the weighting function. There are a number of places in the programs at which measures in interface units must be converted to system values as the system is also being simulated. For instance, the rate constants give rates measured in 1 sec. units, which must be converted to interface time units.

Program II This is the main program of the controller and system simulation. As both controller and system are simulated, the system cannot progress while the controller is making its decisions, which happens in real-life situations. Instead, the simulation is in two parts; one selects the control sequence, as described in Table 3; the other directs the system simulation using these values, checks system performance at the predicted times and, if necessary, returns to the first part to repeat the control value computation, as shown in Table 2. System values are corrected to the nearest interface unit by

$$b\,[\text{integer value of } (a+b/2)/b]$$

where a is the system value and b the interface unit measured in system units.

6. PERFORMANCE

Two different studies were made of the controller's performance.

On-line Study In this study program II was used to control C after a single step change in A. One of these runs is shown in Figure 9 and shows clearly FRED's inefficiency when C is near to the target value, due to the limits on the choice of control value and the turn around parameters. This will be discussed later.

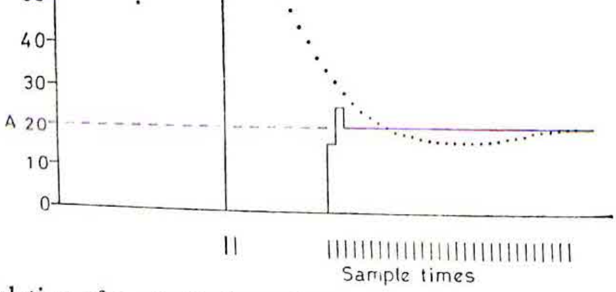


FIGURE 9. Simulation of controller's performance in response to a step change in A.

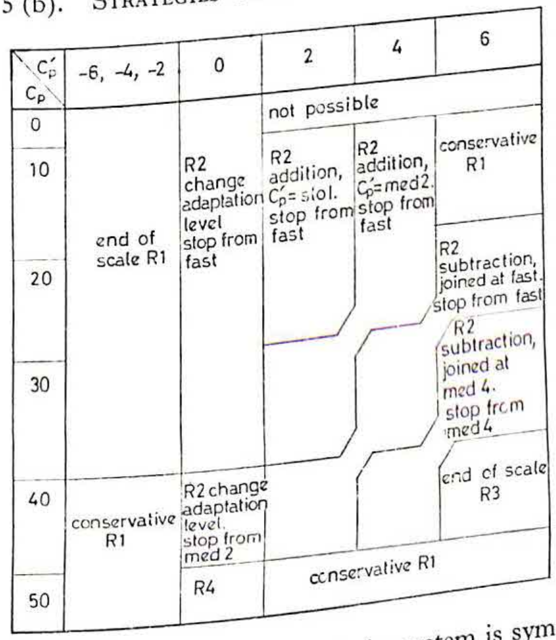
System state/Action state (SS/AS) Study In this study only the control sequence selection part of program II (Table 3) was used. It suggested a control sequence on the basis of a present *system state*, i.e. (C, C') pair, for each cell of a matrix of (C, C') values. The first value of each sequence for $A=50$ is given in Table 5 (a). The underlying strategies are shown in Table 5 (b).

TABLE 5 (a). SYSTEM STATE/ACTION STATE STUDY: FIRST M VALUES SUGGESTED BY FRED

			C_p'				
C_p	−6	−4	−2	0	2	4	6
0	100	100	100	88	—	—	—
10	100	100	100	80	88	88	3
20	100	100	100	73	98	98	93
30	100	100	100	65	68	79	64
40	68	68	57	58	65	63	0
50	66	65	57	50	42	34	33
60	100	37	35	42	40	27	27
70	36	21	32	35	0	0	0
80	7	2	2	27	0	0	0
90	96	12	12	20	0	0	0
100	—	—	—	12	0	0	0

C_p' measured from $t=-t_1$ to $t=0$.

TABLE 5 (b). STRATEGIES UNDERLYING VALUES IN TABLE 5 (a)



C_p \ C_p'	−6, −4, −2	0	2	4	6
0	end of scale R1	R2 change adaptation level stop from fast	not possible		
10			R2 addition, C_p'= stol. stop from fast	R2 addition, C_p'= med2. stop from fast	conservative R1
20					R2 subtraction, joined at fast. stop from fast
30					R2 subtraction, joined at med 4. stop from med 4
40	conservative R1	R2 change adaptation level. stop from med 2			end of scale R3
50		R4	conservative R1		

Only half of the full table is given as the system is symmetrical.

This is parallel to a technique used by Cooke (1965) with human subjects. The system used in his experiments was of the type that FRED can control, a water-bath containing an electric heating coil. The subjects had to change the water temperature to a given value and maintain it there by controlling the power supply to the heating coil. After a number of training sessions with the system, Cooke gave the subjects a sheet marked with a system state matrix and asked them to put in each cell the control setting that they would use if the system was in this state. He called the cells which were completed with the same control setting an *action state*. One of these completed matrices is given in Table 6. The water bath heats more quickly than it cools, while rising and falling are the same in the system controlled by FRED, so care should be taken when comparing the two SS/ASs. FRED could easily be made to control such a system if it contained two sets of system parameters.

TABLE 6. CONTROL SETTINGS CHOSEN BY EXPERIENCED SUBJECT FOR A SYSTEM IN DIFFERENT STATES

System: Water bath with electric heating coil (90° goal) (Cooke, 1965)

	TEMPERATURE RATE OF CHANGE				
TEMPERATURE	*fast down*	*slowly down*	*stationary*	*slowly up*	*fast up*
71					
72					
73					
74					
75					
76					
77					
78					200
79					150
80				200	150
81				150	150
82				150	120
83				150	120
84				150	120
85		200	200	150	120
86		150	150	120	100
87		150	150	120	100
88	200	150	150	120	100
89	150	120	105	100	95
90	150	110	105	100	95
91	110	100	97	95	90
92	100	95	70	50	0
93	100	95	50	0	
94	95	95	0		
95	0	0			
96					
97					
98					
99					

FRED had to assume an M_p value when making its decisions in this task, as the future behaviour of C in the present trajectory is relevant to the control decisions. It assumed that C_p' was the maximum rate of the present trajectory, and that C_p was in the centre of the fast section. Presumably the human subjects made some assumption about the M_p value, but nothing is known about this; their assumption might depend on the C_p' and (C_p to A) values, and this dependence will influence the comparability of human and controller performance.

7. DISCUSSION

The two performance studies raise many similar points.

Controller sensitivity Although a basic feature of this controller is that it has limited ability to distinguish different values of an internal variable, this limited accuracy does not appear consistently in the present version, and should be extended to all variables considered. Only the rate variable has been categorized explicitly; some other variables have been categorized implicitly as their values are obtained from calculations based on categorized values. This is the case with M values, which are chosen by FRED on the basis of C'. As only a few C' values can be envisaged, only a few distances can be used. The other possible values do not 'exist' for the controller, so it shows categorization behaviour, using a number of characteristic values for this variable rather than making an explicit categorization of the continuum.

Although FRED can only use discrete categories, the ratio and adaptation level mechanisms together allow variation in the absolute accuracy of its discriminations, as discussed earlier. This mechanism is an important aspect of the controller as it allows it to change its control accuracy. Choice of R2 value depends on the (C_p to A) distance, as this determines the control trajectory size. This size in turn determines the C' values considered, and the adaptation level, so that judgements based on AL are made more accurately, as C draws nearer to A.

Basing decisions on single characteristic values means that individual M values such as 97, 98, 99 are suggested; it would be more reasonable to use characteristic regions which would suggest a control region on the M scale, of which the characteristic value would actually be used, e.g., 100 in the above cases. The size of these regions would depend on AL and so would be smaller when C_p is near A. This suggestion is a reminder of the complexity required in a digital device designed to simulate a range of values treated as a unit.

Minimum trajectory limits In the present system, t_1 and $k/(1+w)^2$ are so small that z is large, which suggests that the w used here may not be suitable. It might also be necessary to use fewer categories to obtain discriminable slo1 and med2. Cooke (1965) found that his subjects could discriminate five rate values. Alternatively, discriminable values might be obtained with the same w and a much slower system, which it would have been expensive to simulate while debugging the controller program, or a higher order system, in which acceleration takes up a greater proportion of the t template.

The effect of using a constant proportion of the minimum trajectory as a limit on control choices appears both in ($C=40$, $C'=6$) and the mirror cell in the system state table, and in the final part of the on-line run, where it was not possible for FRED to choose an R2 value once C was nearer to A than $z/5$, the limit on R3. Similarly R1 must cause turn around at least $z/2$ from A, 17 units in this system, so conservative R1 values are chosen for C nearer to A than 17 units. These limits were chosen arbitrarily and the controller's performance suggests that a varying limit based on AL would be more efficient.

Turn around parameters The turn around parameters do not give good predictions when compared with actual system behaviour in the on-line runs, especially at extreme values. This is the cause of the entry in ($C=10$, $C'=6$), where the parameters predicted that C could not be stopped without overshooting A, so a conservative R1 was chosen to stop C at $z/2$ beyond A.

When FRED is testing R2 and R3 values while choosing a control sequence, if the turn around parameters are smaller than those which correctly predict system performance, R2 values will be accepted that will actually cause C to overshoot A; while if they are larger, C will not have reached A when FRED looks at the effects of control. In both cases the approach to A will be inefficient. FRED also uses the turn around parameters during control, when checking the effect of using R2 (note 1 to Table 2). In this case, if the parameter is too small, FRED will make frequent samples and recalculations of its strategy; while if it is too large, C will always overshoot A.

A single parameter is obviously an oversimplified description of real system behaviour, but if the controller is modelling methods of prediction rather than attempting to make perfect predictions, this is not a relevant criticism. The method chosen has a great influence on the effectiveness of FRED's decisions and the pattern of M values in the SS/AS. The controller could make turn around predictions in other ways. For instance, it could assume a single value for turn around distance and time, independent of the forward and reverse distances, or choose R1 and R3 values which would give a standard turn around time rather than a standard turn around distance, or use an R1 value which gave a standard turn around distance rather than a standard distance of C_T from A.

The response to different rates at the same C value In the human SS/AS different rates are controlled by different suggested M values which change monotonically across the table. In the controller, because of the nested nature of rate categorization, approximately the same M values are suggested in response to different rates on the basis of different strategies. This in combination with uncategorized M values gives rather irregular numbers in Table 5 (a), although Table 5 (b) shows that the strategies change in a regular manner. ALs associated with these strategies are different, so that some conservatism based on adaptation level might give a more realistic performance. This similarity of responses might also depend in part on the turn around parameters, and on differences in M_p values assumed by the controller and the human subject.

The response to different C values at the same rate Controller values are again not monotonic, although strategies change regularly. An exception is the relation between M and C_p for $C' = 0$, which is linear. In the human SS/AS this relation is sigmoidal, with greater sensitivity to M values as C is nearer the target value. Again, more realistic behaviour might be obtained from a controller with some conservatism depending on adaptation level. However, the success of FRED as a model for human performance would then depend on the modelling of conservatism rather than on the strategies used in its present form.

8. Conclusions

The purpose of this work was to develop a controller with some human characteristics, and thereby to investigate these characteristics and show that they imply others. The basic feature of the controller is a device which can only distinguish between a limited number of values at one time but can discriminate at any level of accuracy if these values are determined by ratios and an adaptation level; if w and AL are involved in the processes of control decision then control accuracy is necessarily similarly affected. It would be useful to replace FRED's more arbitrary features by data on human performance. For instance, do human controllers show this nested type of discrimination for continuous variables, and the limitations on prediction which follow from it? Do human controllers use simple categorized templates to make predictions? And if so, in what way do they use them; and do the templates determine the timing of their samples and corrective actions? This paper demonstrates that a device of the kind described can control a system, and so shows that it would be meaningful to investigate human control behaviour in terms of such a model.

Acknowledgements

The model was simulated in FORTRAN IV on an IBM 7090/94 at the Computer Center, University of California, Berkeley. This work was supported in part by USPHS AC-260-01 Crossman.

The author would like to thank Prof. E. R. F. W. Crossman of the Department of Industrial Engineering, University of California, Berkeley, for his interest and encouragement, and Dr. R. J. Beishon of the Psychology Department, Bristol University, for his helpful comments on the manuscript.

References

Beishon, R. J. (1966). *A Study of Some Aspects of Mental Skill in the Performance of Laboratory and Industrial Tasks.* Unpublished D.Phil. Thesis, Oxford University.

Beishon, R. J. (1967). Problems of task description in process control. *Ergonomics*, **10**, 177–186.

COOKE, J. E. (1965). *Human Decisions in the Control of a Slow-Response System.* Unpublished D.Phil. Thesis, Oxford University.
McCORMICK, E. J. (1964). *Human Factors Engineering.* New York: McGraw-Hill.
MILLER, G. A. (1956). The magical number seven, plus or minus two; some limits on our capacity for processing information. *Psychol. Rev.* **63**, 81–97.
POULTON, E. C. (1957). On prediction in skilled movements. *Psychol. Bull.* **54**, 467–78.
SENDERS, J. W. (1964). The human operator as a monitor and controller of multi-degree-of-freedom systems. *Trans. I.E.E.E.* HFE–5, September, 1–6.
SHERIDAN, T. B. (1962). The Human Operator in Control Instrumentation. In *Progress in Control Engineering.* London: Heywood & Co.

Vol. 20 The British Journal of Mathematical and May
Part 1 Statistical Psychology 1967
51–62

TYPE I ERRORS AND THEIR DECISION BASIS

By R. S. Rodger
University of Sydney

It is proposed that an error-rate I should be set for each statistical decision to be made. Methods for testing statistical hypotheses in the context of analysis of variance are reviewed and evaluated from the decision-based error-rate viewpoint. If decisions are the proper bases for the choice of an error-rate, the available methods for testing any or all linear contrasts or linear functions of the means are unacceptable. Two new methods (R and D) are proposed for testing all contrasts or linear functions. One implication of the arguments given here is that, for most purposes, research workers have been using the wrong tables of F to test their statistical hypotheses.

1. Introduction

Ryan (1959) brought to the notice of psychologists the question of the basis for the choice of an error-rate. He argued that the experiment should be accepted as the basis. A single error-rate should be chosen for each experiment, this being the probability of making at least one false decision. But the definition of an experiment is usually arbitrary and the experimental limits variable from occasion to occasion. These are practical grounds for questioning the unqualified acceptance of experiment-based error-rates. Theoretically, the use of the experiment as the basis for the choice of an error-rate seems arbitrary and unnatural. The bases for choosing error-rates must surely be the occasions when errors may be made—decisions.

This point of view applies to testing all types of statistical hypotheses and three cases are distinguished later in the paper—testing a prescribed set of hypotheses, goodness-of-fit tests and testing an unlimited number of hypotheses. These methods are elaborated solely in terms of one-way, univariate analyses of variance (with fixed treatments), but the basic idea of decision-based error-rates applies in all statistical contexts, in multi-classification designs and multivariate models, in the analysis of attributes by their frequency of occurrence as well as in the analyses of normal variates.

It will be assumed that there are a treatment groups or populations with means μ_j and sample means M_j $(j=1, 2, \ldots, a)$. Samples will be considered to be random with equal replication sizes n, and to have normal error.

2. Contrasts

The most frequently tested linear statistical hypothesis about the means of normal variates can be expressed as a contrast. The general form for the ith contrast is

$$\kappa_i = c_{i1}\mu_1 + c_{i2}\mu_2 + \ldots + c_{ia}\mu_a = 0. \tag{1}$$

The coefficients c_{ij} are chosen by the investigator because of their relevance to the purpose of his study. The restrictions on the choice of the c_{ij}'s are that they be real numbers, not all zero, such that $\sum_j c_{ij}=0$.

If, for fixed i, one value of $c_{ij}=1$, another $c_{ij}=-1$ and all other $c_{ij}=0$, the contrast is a simple difference between two means and is termed a comparison. The contrasts

$$\kappa_1=\mu_1-\mu_2=0; \quad \kappa_2=0{\cdot}5(\mu_1+\mu_2)-\mu_3=0;$$

and

$$\kappa_3=\mu_1+\mu_2+\mu_3-3\mu_4=0;$$

for $a=4$, have contrast coefficients and sums of squares

$$c_{1j}=1,\ -1,\ 0,\ 0 \text{ and } \sum_j c_{1j}^2=2; \quad c_{2j}=0{\cdot}5,\ 0{\cdot}5,\ -1,\ 0 \text{ and } \sum_j c_{2j}^2=1{\cdot}5$$

and

$$c_{3j}=1,\ 1,\ 1,\ -3 \text{ and } \sum_j c_{3j}^2=12.$$

A component of variation is a second-degree expression which is equivalent to a contrast, since a test of the component will always yield the same decision as a test of its corresponding contrast. The ith component is

$$\Lambda_i=n(\sum_j c_{ij}\mu_j)^2/\sum_j c_{ij}^2=n\kappa_i^2/\sum_j c_{ij}^2=0. \tag{2}$$

If several contrasts are specified for the same population means, the relations between them can be described by their intercorrelations. The correlation between any two contrasts κ_1 and κ_2 is

$$\rho_{12}=\sum_j c_{1j}c_{2j}/(\sum_j c_{1j}^2 \sum_j c_{2j}^2)^{\frac{1}{2}}. \tag{3}$$

When $\rho_{12}=0$ the contrasts are orthogonal. When $\rho_{12}^2=1$, the contrasts will be termed identical contrasts because their components of variation are identical. When $0<\rho_{12}^2<1$, the contrasts are similar though different. The three contrasts quoted above are mutually orthogonal. Contrast κ_2 is identical to $\mu_1+\mu_2-2\mu_3=0$ and κ_2 is similar to $\mu_1-\mu_3=0$, their correlation being $+0{\cdot}866$. For any contrast κ_i we may choose any real number w_i and write $w_i\kappa_i$ to obtain an identical contrast.

When $a>2$ an infinite number of different linear contrasts are possible among the a means. If the contrast coefficients for p contrasts are written as the elements of a matrix ${}_p\mathbf{C}_a$, the rank of this matrix cannot exceed $(a-1)$, no matter what the size of p. This indicates that any contrast can be expressed as the weighted sum of a set of $(a-1)$ contrasts, the coefficient matrix of which has rank $(a-1)$. It is for this reason that there are $(a-1)$ degrees of freedom available for contrasts among a means. The sample estimators of contrasts and components of variation are

$$k_i=c_{i1}M_1+c_{i2}M_2+\ \dots\ +c_{ia}M_a; \tag{4}$$

and

$$L_i=n(\sum_j c_{ij}M_j)^2/\sum_j c_{ij}^2=nk_i^2/\sum_j c_{ij}^2. \tag{5}$$

The component estimators corresponding to κ_1 and κ_2 above are

$$L_1 = n(M_1 - M_2)^2/2$$

and

$$L_2 = n[0{\cdot}5(M_1 + M_2) - M_3]^2/1{\cdot}5 = [(M_1 + M_2)/2 - M_3]^2/[(1/2n) + (1/n)].$$

If each of these were divided by the estimate of error variance, they would be identical to the squares of t tests.

If $p = (a-1)$ contrasts are specified, with coefficient matrix rank $(a-1)$, the corrected sum of squares between sample means (SSB) can be expressed in terms of these p contrasts. Let ${}_p\mathbf{k}_1$ be a vector, the elements of which are the sample values of the p contrasts. Then

$$\text{SSB} = n_1\mathbf{k}_p'({}_p\mathbf{C}_a\,{}_a\mathbf{C}'_p)^{-1}\,{}_p\mathbf{k}_1. \tag{6}$$

If the p contrasts are mutually orthogonal, $({}_p\mathbf{C}_a\,{}_a\mathbf{C}'_p)$ is a diagonal and

$$\text{SSB} = \sum_{i=1}^{p} L_i. \tag{7}$$

Contrast coefficients may be chosen by the investigator either independently of the data used for testing purposes or dependent upon them. Which course the investigator chooses partly determines the appropriate sampling distribution for testing the contrasts. An interesting example of the second approach is the maximized, observed contrast, the coefficients of which, for $a > 2$, can be determined only from the data. The coefficients for the maximized contrast (κ_m) are

$$c_{m1} = M_1 - M_{\cdot},\ c_{m2} = M_2 - M_{\cdot},\ \ldots\ c_{ma} = M_a - M_{\cdot},$$

where $M_{\cdot}$ is the mean of all samples combined. For this contrast

$$L_m = n\sum_j (M_j - M_{\cdot})^2 = n(\sum_j c_{mj}M_j)^2/\sum_j c_{mj}^2 = \text{SSB}. \tag{8}$$

The basic error term for contrasts is the mean-square within treatments (MSW) in analysis of variance, which is an estimator of the population error variance σ_ϵ^2 and has $\nu_2 = a(n-1)$ degrees of freedom. A number of different sampling distributions may be used to test contrasts and these will not generally give identical decisions. The investigator must choose, independently of his data, which sampling distribution will be used. The emphasis in this paper will be on the variance ratio distribution and the studentized range distribution for testing contrasts. The general form of all tests may be written

$$F = L_i/\nu_1\text{MSW} \tag{9}$$

or

$$q^2 = 4nk_i^2/(\sum_j |c_{ij}|)^2\text{MSW} \tag{10}$$

where q is the studentized range distribution for range $\nu_1 + 1$, with error degrees of freedom ν_2, and F is the variance ratio distribution with ν_1 and ν_2 degrees of freedom.

Linear statistical hypotheses which are not contrasts are sometimes tested. These are also of the form given by eqn. (1) but with $\sum c_{ij} \neq 0$. Thus, for

$a=4$, it may be asserted that the average $\mu_j=0$ by choosing coefficients $c_{0j}=1, 1, 1, 1$ and $\sum_j c_{0j}^2=4$. Such hypotheses may be treated in exactly the same way as contrasts, as described above. This particular hypothesis is orthogonal to each of the three contrasts specified earlier, as substitution of the coefficients in eqn. (3) will show. Another of many possible orthogonal sets is

$$ {}_4\mathbf{C}_4=\begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix}. $$

A matrix of coefficients ${}_p\mathbf{C}_a$, which contains at least one row with $\sum_j c_{ij}\neq 0$, cannot have a rank greater than a, no matter what the size of p. Such a matrix with rank a has a degrees of freedom available for tests carried out upon its corresponding components.

If the uncorrected sum of squares between observed means (USSB) is used rather than the corrected sum of squares (SSB), a matrix ${}_p\mathbf{C}_a$ with $p=a$ and rank a will give

$$\mathrm{USSB}=n_1\mathbf{k}_p'({}_p\mathbf{C}_a\,{}_a\mathbf{C}_p')^{-1}{}_p\mathbf{k}_1, \tag{11}$$

which is the rank a version of eqn. (6). If the a components are mutually orthogonal then

$$\mathrm{USSB}=\sum_{i=1}^{p} L_i \tag{12}$$

which is the rank a version of eqn. (7).

There is no absolute requirement, either for contrasts or other linear hypotheses, that the null case be specified or tested. The most general linear hypothesis is

$$\kappa_1=c_{i1}\mu_1+c_{i2}\mu_2+\ldots+c_{ia}\mu_a+d_i=0,$$

and d_i need not be zero. The sample estimator for the component which corresponds to this expression is

$$L_i=n(\sum_j c_{ij}M_j+d_i)^2/\sum_j c_{ij}^2. \tag{13}$$

The general form of the test for such a component is still eqn. (9). The sample estimator of κ_i is

$$k_i=c_{i1}M_1+c_{i2}M_2+\ldots+c_{ia}M_a+d_i, \tag{14}$$

and this may be used in eqn. (10).

3. PRESCRIBED TESTS

The statistically most powerful way to test contrasts or other linear components is to select a or fewer of them independently of the date used to test them, hence preferably before such data are collected. If $p \leqslant a$ are selected, their coefficient matrix having rank p, the components of variation for each may be inserted in eqn. (9), with $\nu_1 = 1$, and each component rejected for which the resulting $F \geqslant F_{1-\alpha}$ from the standard tables, with $\nu_1 = 1$. These tests are exactly equivalent to t tests and α is the type I error-rate for each decision made. This strategy uses the shortest possible confidence intervals for κ_i, and treats the experiment as p sub-experiments, designed as a unit for greater efficiency. This is a method recommended by Fisher (1951, pp. 198–203).

If a mutually orthogonal components are selected and the probability of rejecting each in error set at α, the probability that each will be retained correctly is $(1-\alpha)$. These a probabilities are statistically independent; so the probability of retaining all components correctly is $(1-\alpha)^a$. The probability of not retaining all components correctly, i.e. rejecting at least one in error is $1-(1-\alpha)^a$. This is the experiment-based error-rate. If the selected components are not mutually orthogonal, the probabilities $(1-\alpha)$ are not statistically independent and the experiment-based error-rate may be less than $1-(1-\alpha)^a$. Even for mutually orthogonal contrasts the decisions that are made will not be mutually independent if all tests are performed with a common MSW. Such dependence may be removed by splitting the sum of squares within treatments into a parts, at random, and using each of these parts to provide independent estimates of σ_ϵ^2, each with $(n-1)$ degrees of freedom. Each of the a tests will now use a different MSW. In practice the use of a common MSW seems to have a negligible effect on the distribution of decision errors.

It is true that as a increases, the probability of rejecting at least one component in error also increases: but this is equally true if we adopt the much less efficient scheme of keeping a small and increasing the number of experiments.

If only contrasts are to be tested by the methods of this section, then a should be replaced by $(a-1)$ in the last two paragraphs.

As an alternative to inserting estimated components in eqn. (9), the estimates k_i may be inserted in eqn. (10). If the square-roots of the resulting values of q^2 are extracted, each will have the studentized range distribution for a range of two with $\nu_2 = a(n-1)$ degrees of freedom for error. Each contrast or linear hypothesis for which the observed $q \geqslant q_{1-\alpha}$ should be rejected. Here α is the type I error-rate for each decision to be made. This test will give identical decisions to those arising from the use of eqn. (9).

4. GOODNESS OF FIT

It occasionally happens that an experiment is designed to test the goodness-of-fit of observed data to some theoretical statement. A simple example is the hypothesis that all $\mu_j = 0$. This may be tested by setting USSB $= L$ and inserting this in eqn. (9), with $\nu_1 = a$. If the resulting $F \geqslant F_{1-\alpha}$ for $\nu_1 = a$ and $\nu_2 = a(n-1)$, the hypothesis is rejected. The type I error-rate for this decision is α. This is a strictly one-decision procedure. Should further tests on the same data be intended if the null hypothesis is rejected, this method is not appropriate and the procedures of Section 5 should be adopted.

Another hypothesis which may be tested by the method of this section is that all μ_j are equal. This is tested by inserting the observed value of L_m in eqn. (9) with $\nu_1=a-1$ and rejecting the hypothesis if the resulting $F\geqslant F_{1-\alpha}$ for $\nu_1=a-1$ and $\nu_2=a(n-1)$. This is again a strictly one-decision process with α being the type I error-rate for that decision.

In spite of the fact that the last-mentioned hypothesis (equivalent to $\Lambda_m=0$) is very frequently tested, the view taken here is that it is only rarely an hypothesis of primary interest in empirical research. The evidence for this view is the high frequency with which predictable contrasts are tested in data for which $\Lambda_m=0$ has been rejected. If $\Lambda_m=0$ is true, then all $\Lambda_i=0$ are true. Conversely, if all $\Lambda_i=0$ are true, then $\Lambda_m=0$ is true. These two directions of logical argument have critically different statistical properties and it is essential to specify which terms are the premises and which the conclusions in an investigation before devising the appropriate statistical tests. When the $\Lambda_i=0$ are primary, as they usually are, the methods of Sections 3 or 5 are appropriate for testing them.

The goodness-of-fit strategy is appropriate in some curve-fitting applications. In such studies, levels of the treatment variable $(X_1, X_2, \ldots, X_a)$ occur in an hypothetical equation which relates them to the population means (μ_j). An example is the linear regression equation

$$\mu_j=g_0+g_1X_j. \tag{15}$$

If we write $c_{tj}=X_j-\bar{X}$, where $\bar{X}$ is the mean of the X's, the c_{tj} are real numbers, not all zero, and $\sum_j c_{tj}=0$. The c_{tj} are the contrast coefficients for the linear trend component

$$\Lambda_t=n(\sum_j c_{tj}\mu_j)^2/\sum_j c_{tj}{}^2=0. \tag{16}$$

If the true relation between the μ_j and X_j is given by eqn. (15), the variation of the M_j around the trend line will be random and normal with an expected mean of zero. The observed amount of variation of the M_j around the trend line may be measured by

$$L_r=L_m-L_t. \tag{17}$$

If eqn. (15) is the only true one, then $\Lambda_t=0$ is false and $\Lambda_r=0$ is true. Should the former be retained or the latter rejected, eqn. (15) must be discarded.

$\Lambda_t=0$ may be tested by the methods of Section 3. We test $\Lambda_r=0$ by inserting L_r in eqn. (9) with $\nu_1=(a-2)$ and reject it if the resulting $F\geqslant F_{1-\alpha}$, from the standard tables, with the stated ν_1 and ν_2. Once more α is the type I error-rate for this decision and further tests on the same data are not carried out, no matter which decision is made.

5. UNRESTRICTED TESTS

It frequently happens that investigators wish to test more than $(a-1)$ contrasts, or more than a components at least one of which has $\sum_j c_{ij} \neq 0$. It is also common for investigators to seek to reserve the right to search among their obtained data to find relations among the μ_j which can be interpreted. Exploring data can lead to abuses, one of which is the testing of sets of retained and rejected contrasts, with no systematic attempt at empirical rationalization or interpretation, and another is exaggerated claims to logical status for empirical interpretations placed on the decision set without adequate support from prior or subsequent research. These abuses, though serious, are not especially statistical problems. The principal statistical problem is to provide a method for testing more than the prescribed number of components specified in Section 3, or for testing any components suggested by a scrutiny of the data with known error-rates α.

Whenever p components are tested, the coefficient matrix of which has rank less than p, there exists the possibility that intransitive decisions may be made. An example of intransitive decisions is

$$\mu_1-\mu_2=0; \quad \mu_2-\mu_3=0 \quad \text{and} \quad \mu_1-\mu_3\neq 0.$$

If the first two decisions are made, then $\mu_1=\mu_2=\mu_3$ and $\mu_1=\mu_3$. The methods reviewed, and the new methods introduced in this section do not prevent such intransitivity happening. It is left to the investigator to choose the decision set or sets he will interpret.

For many years it was the practice to insert L_m in eqn. (9) with $\nu_1=a-1$, and, if $\Lambda_m=0$ was rejected, insert the various L_i of interest in eqn. (9) with $\nu_1=1$. This is the method of protected t tests. This procedure has been known to be unsatisfactory for over 30 years. Among other things, it will yield a distribution of decision errors which are not random in Λ_i; the decision errors being concentrated in the experiments in which $\Lambda_m=0$ is rejected, and being greater than 100α per cent in these experiments.

Scheffé (1953) suggested that any component of interest should be tested by inserting its L_i in eqn. (9) with $\nu_1=(a-1)$, and rejecting $\Lambda_i=0$ if the computed $F \geqslant F_{1-\alpha}$. This method uses the experiment as the basis for its error-rate choice, rather than the multiple decisions which will be made. The experiment as the basis for error-rate choice is traditional but in the author's opinion incorrect where multiple decisions are to be made.

It is proposed here that the Scheffé procedure be modified simply by using tables of $F_{(1-\alpha)^{\nu_1}}$ to test the F of eqn. (9) instead of the traditional tables of $F_{1-\alpha}$. Tables of $F_{(1-\alpha)^{\nu_1}}$ have been provided by Rodger (1965).

The justification for this new procedure is that $(1-\alpha)^{\nu_1}$ is the *minimum* probability of correctly accepting all contrasts, or other linear components, each set at $(1-\alpha)$. If two components are tested with type I error-rate α, the probability that both will be correctly accepted is the probability of accepting one, $(1-\alpha)$, multiplied by the conditional probability of accepting the other.

When the two components are orthogonal, the conditional probability of accepting the second correctly, given the correct acceptance of the first, is $(1-\alpha)$ and the probability of accepting both correctly is $(1-\alpha)^2$. When the two components are identical, such as κ_1 and $\kappa_2 = w_1\kappa_1$, the conditional probability of correctly accepting the second is not $(1-\alpha)$ but 1, and the probability of correct acceptance of both is not $(1-\alpha)^2$ but $(1-\alpha)$. Any component in excess of a set of ν_1, with coefficient matrix rank ν_1, is perfectly predictable from that set in the form.

$$\kappa_i = w_1\kappa_1 + w_2\kappa_2 + \ldots + w_{\nu_1}\kappa_{\nu_1}.$$

The conditional probability of accepting any such component, given the acceptance of the ν_1 others, is therefore 1. The probability of accepting correctly all components is the product of a sequence of probabilities, those beyond the ν_1th being 1. This product has a minimum value $(1-\alpha)^{\nu_1}$ when the set of ν_1 consists of mutually orthogonal components. The *maximum* probability of rejecting at least one component in error ($\Lambda_m = 0$ in the contrast case) is thus $1-(1-\alpha)^{\nu_1}$. The probability of rejecting $\Lambda_m = 0$ in error will be $1-(1-\alpha)^{\nu_1}$, if we insert $L_i = L_m$ in eqn. (9) and evaluate this against $F_{(1-\alpha)^{\nu_1}}$. The sampling distribution of L_m is F with $\nu_1 = (a-1)$ and $\nu_2 = a(n-1)$ degrees of freedom—as is well known. With a strategy of making unrestricted tests there is no need to test $\Lambda_m = 0$, a component of little empirical interest, but testing it will generally save time, since if $\Lambda_m = 0$ is accepted all other observed values of F from eqn. (9) will be less than $F_{(1-\alpha)^{\nu_1}}$. This may be used to test components which are not contrasts by setting $\nu_1 = a$.

Scheffé defined his test criterion as $S = (\nu_1 F_{1-\alpha})^{\frac{1}{2}}$. We shall redefine this to be

$$S_i = (\nu_1 F_{1-\alpha} \sum_j c_{ij}^2 \,\mathrm{MSW}/n)^{\frac{1}{2}} \tag{18}$$

which is the size of the confidence interval for a contrast, or other linear expression k_i. This is a useful expression which may be calculated once, then used to search among the observed values of k_i which have a common $\sum_j c_{ij}^2$. The new method introduced here can also be expressed in terms of a confidence interval length, which is

$$R_i = (\nu_1 F_{(1-\alpha)^{\nu_1}} \sum_j c_{ij}^2 \,\mathrm{MSW}/n)^{\frac{1}{2}}. \tag{19}$$

The decision rule can be reformulated in terms of this quantity as,

$$\text{Reject } \kappa_i \text{ if } |k_i| \geq R_i. \tag{20}$$

If we employ this rule the probability that at least one contrast (or other linear component) will be rejected in error *cannot exceed* $1-(1-\alpha)^{\nu_1}$.

Duncan (1955) suggested using $F_{(1-\alpha)^{\nu_1}}$, but he probably intended that this be used to test all comparisons in a nested fashion in place of his DMR method, which will be discussed later.

Tukey (1953) suggested using the studentized range distribution as the criterion for unrestricted tests. This method inserts the observed value of any contrast k_i in eqn. (10) and rejects κ_i if the resulting $q \geqslant q_{1-\alpha}$ which is the studentized range value for a range of a. If we modify Scheffé's notation once more, and define

$$T_i = \tfrac{1}{2}\sum_j |c_{ij}| q_{1-\alpha} (\mathrm{MSW}/n)^{\frac{1}{2}}, \tag{21}$$

the decision rule proposed by Tukey is

$$\text{Reject } \kappa_i \text{ if } |k_i| \geqslant T_i. \tag{22}$$

As was true of the S technique, the T method uses the experiment as the basis for error-rate choice. The method can be modified to give the error-rate a decision basis in the case where unrestricted tests are to be made by using

$$D_i = \tfrac{1}{2}\sum_j |c_{ij}| q_{(1-\alpha)^{\nu_1}} (\mathrm{MSW}/n)^{\frac{1}{2}} \tag{23}$$

and the decision rule

$$\text{Reject } \kappa_i \text{ if } |k_i| \geqslant D_i. \tag{24}$$

Duncan (1951) uses $q_{(1-\alpha)^{\nu_1}}$ in a test procedure (DMR) for unrestricted comparisons, but this is not identical to the D method proposed above. Extensive tables of $q_{1-\alpha}$ and $q_{(1-\alpha)^{\nu_1}}$ have been produced by Harter *et al.* (1959), an abridged version of $q_{1-\alpha}$ by Harter (1960 a) and of $q_{(1-\alpha)^{\nu}}$ by Harter (1960 b). Abridged tables of both test values are given in Rodger (1965).

D and T are simple range methods, but multiple-range methods have also been proposed to solve the unrestricted tests problem. Newman (1939) and Keuls (1952) have developed a procedure first mentioned by Student in a letter he wrote to E. S. Pearson in 1932 (see Pearson, 1938). This technique arranges the observed means in increasing order of size, say $M_1, M_2, \ldots, M_a$, and tests comparisons between them from the outermost inwards, in a nested fashion. The interval

$$\mathrm{KMR}_h = q_{1-\alpha} (\mathrm{MSW}/n)^{\frac{1}{2}} \tag{25}$$

is set up with $q_{1-\alpha}$ being the critical studentized range statistic for a range of $h+1$ means, with error degrees of freedom $\nu_2 = a(n-1)$. The outermost comparison $(\mu_a - \mu_1 = 0)$ is rejected if $(M_a - M_1)$ reaches or exceeds KMR_h for $h = a-1$. If this comparison is retained, all are retained. If it is rejected, the two comparisons inside it ($\mu_a - \mu_2 = 0$ and $\mu_{a-1} - \mu_1 = 0$) are tested against KMR_h for $h = a-2$. If either or both of these are rejected, the comparisons inside either or both are then tested against KMR_h with $h = a-3$. The process terminates when all comparisons for fixed h are retained.

KMR uses a variable basis for error-rate choice, the range a being first, then $a-1$, $a-2, \ldots 2$. The first of these uses an experiment rather than a decision as a basis. Duncan (1951) modified this to a decision basis by using

$$\mathrm{DMR}_h = q_{(1-\alpha)^h} (\mathrm{MSW}/n)^{\frac{1}{2}} \tag{26}$$

to reject comparisons. This is still a multiple-range test and uses a variable basis for error-rate choice—$a-1$ decisions, $a-2$ decisions, $a-3$ decisions . . . 1 decision, yet the number of decisions made directly at each level is the reverse of this, i.e., 1, 2, . . . $a-3$, $a-2$ and $a-1$.

Multiple-range methods are for comparisons and may not be used for other types of contrasts. They have some affinity with the protected t test, not only in the modified basis for error-rate choice but also because of the rule that no comparison may be rejected, even though its $|k_i| \geqslant \mathrm{KMR}_h$, if it is contained within another which is retained.

The methods for making unrestricted numbers of tests recommended here are either R or D. These two methods will not always reject the same components and the investigator must choose which he will use independently of the data.

The relation between these and other methods may be elucidated by comparing the lengths of the confidence intervals for κ_i which each uses. These lengths are shown for comparisons with $a=5$ and $a=10$ in Table 1. The multiple-range methods use varying interval lengths but, if we use their average, weighted lengths, the increasing order of length is t, DMR, D, KMR, R, T and S. When contrasts other than comparisons are used, the order of lengths changes. This can be seen in Table 2 which gives the lengths when $a=5$ and $a=10$ for the contrasts

$$\kappa_1 = \mu_1 + \mu_2 - 2\mu_3 = 0$$

and

$$\kappa_2 = 3\mu_1 + 3\mu_2 - 2\mu_3 - 2\mu_4 - 2_5 = 0.$$

The order now is, roughly, t, R, D, S, and T, depending to some extent on a and the contrast being tested.

The differences in confidence interval length between the techniques in Table 1 or Table 2 are due to the different sampling distributions they employ and how the decision error-rates are assigned. For S the probability is $\alpha = 0{\cdot}05$ that the maximized contrast will be rejected in error and T assigns this error-rate to the maximum observed comparison. R and D assign $1-(1-0{\cdot}05)^{\nu_1}$ as the *maximum* probability of false rejection for at least one contrast (R), or comparison (D). DMR and KMR use a variable basis for error-rate assignment.

6. SUMMARY

It has been proposed that a type I error-rate α be set for each statistical decision to be made. In the simple analysis of variance any linear component among the a means ($\Lambda_i = 0$) may be tested by $F = L_i/\nu_1\mathrm{MSW}$. If $p \leqslant a$ components are predicted, with coefficient matrix rank p, the test formula has $\nu_1 = 1$ and we reject the component if $F \geqslant F_{1-\alpha}$ for the degrees of freedom ν_1 and $\nu_2 = a(n-1)$. This is the t test procedure and α is the type I error-rate for each of the p decisions. When a component with p degrees of freedom is tested (such as $\Lambda_m = 0$), and no resolution of this component into parts is intended, the test formula has $\nu_1 = p$ and we reject the component of $F \geqslant F_{1-\alpha}$ for degrees of freedom ν_1 and ν_2. This is a goodness-of-fit, single-decision procedure, the type I error-rate for this being α.

TABLE 1. CONFIDENCE INTERVAL LENGTHS FOR COMPARISONS WITH MSW=9, $n=3$ AND $\alpha=0{\cdot}05$

Number of treatments (a)	5				10								
Number of means compared (h)	1	2	3	4	1	2	3	4	5	6	7	8	9
DMR_h	5·46	5·70	5·85	5·94	5·11	5·36	5·53	5·64	5·72	5·78	5·83	5·87	5·90
KMR_h	5·46	6·72	7·49	8·06	5·11	6·20	6·86	7·33	7·70	8·00	8·26	8·48	8·67
For all h	S=9·14 R=6·77 t=5·46			T=8·06 D=5·94	S=11·37 D=5·90			T=8·67 t=5·11		R=7·91			

TABLE 2. CONFIDENCE INTERVAL LENGTHS FOR THE CONTRASTS

$\kappa_1=\mu_1+\mu_2-2\mu_3=0$ AND $\kappa_2=3\mu_1+3\mu_2-2\mu_3-2\mu_4-2\mu_5=0$; WITH MSW=9, $n=3$ AND $\alpha=0{\cdot}05$

	$a=5$					$a=10$				
	T	S	D	R	t	T	S	D	R	t
κ_1	16·12	15·83	11·88	11·73	9·45	17·35	19·70	11·81	13·71	8·85
κ_2	48·37	35·40	35·65	26·22	21·13	52·05	44·03	35·43	30·65	19·79

When more than a linear components (or $a-1$ contrasts) are to be tested, we set $\nu_1 = a$ (or $\nu_1 = a-1$ for contrasts only) and reject each component for which $F > F_{(1-\alpha)\nu_1}$, with degrees of freedom ν_1 and ν_2. When more than ν_1 decisions are made, the set may contain intransitivities. Such intransitivities may be removed by reducing the set (*post-hoc*) to a set of ν_1, the coefficient matrix of which has rank ν_1. In particular, a set of ν_1 mutually orthogonal components will remove such intransitivities.

The studentized range distribution may be used for any of the above procedures. This distribution will yield identical decisions to F in the first strategy but not, generally, in the other two. The investigator must select, independently of his test data, which distribution he will use.

REFERENCES

DUNCAN, D. B. (1951). A significance test for differences between ranked treatments in the analysis of variance. *Virginia J. Sci.* **2** (New Series), 171–189.

DUNCAN, D. B. (1955). Multiple range and multiple F tests. *Biometrics*, **11**, 1–42.

FISHER, R. A. (1951). *The Design of Experiments* (6th Ed.). Edinburgh: Oliver & Boyd.

HARTER, H. L., CLEMM, D. S. and GUTHRIE, E. H. (1959). *The Probability Integrals of the Range and of the Studentized Range.* Vol. II, W.A.D.C. Tech. Rep. No. 58–484.

HARTER, H. L. (1960 a). Tables of range and studentized range. *Ann. math. Statist.* **31**, 1122–1147.

HARTER, H. L. (1960 b). Critical values for Duncan's new multiple range test. *Biometrics*, **16**, 671–685.

KEULS, M. (1952). The use of studentized range in connection with an analysis of variance. *Euphytica*, **1**, 112–122.

NEWMAN, D. (1939). The distribution of range in samples from a normal population, expressed in terms of an independent estimate of standard deviation. *Biometrika*, **31**, 20–30.

PEARSON, E. S. (1938). Student as statistician. *Biometrika*, **30**, 210–250.

RODGER, R. S. (1965). *Intermediate Statistics.* Sydney: University Bookshop.

RYAN, T. A. (1959). Multiple comparisons in psychological research. *Psychol. Bull.* **56**, 26–47.

SCHEFFÉ, H. (1953). A method for judging all contrasts in the analysis of variance. *Biometrika*, **40**, 87–104.

TUKEY, J. W. (1953). *The Problem of Multiple Comparisons.* Princeton: Author.

Vol. 20
Part 1
63–69

The British Journal of Mathematical and Statistical Psychology

May
1967

THE ASSUMPTION OF MULTIVARIATE NORMALITY

By R. S. Lockhart[1]

University of Sydney

Although the multivariate normal assumption is common to many of the multivariate statistical methods used to analyse psychological data, its tenability is rarely examined. It is possible to derive a goodness-of-fit test of this assumption, but in practice sample sizes will not be large enough to make complete testing feasible. In such a situation a partial testing procedure is advocated, together with studies of robustness. A numerical example is given illustrating how the partial testing procedure may detect instances of distributions which have normally distributed marginals but which are not multivariate normal.

1. Introduction

The assumption of multivariate normality is common to many of the multivariate statistical methods applied to psychological data. Despite its ubiquity there is no method in common use whereby such an assumption can be tested; the consideration of whether an observed set of data conforms to the multinormal model is usually limited to an examination of the normality of the marginal distributions. It is well known, however, that marginal normality is not a sufficient condition for *joint* normality, and it is an easy matter to construct distributions which have normally distributed marginals but which are not multivariate normal (Carroll, 1961, p. 360). More generally, while it is true that if an n-variate distribution is joint normai any subset of the n variates will have a multivariate normal distribution, the converse does not necessarily hold. In particular, the fact that the $n(n-1)/2$ bivariate distributions of an n-variate distribution are bivariate normal does not ensure that the distribution is n-variate normal. An examination of bivariate scatter-plots is therefore of limited value.

As Binder (1959) has pointed out in the context of the bivariate normal model for correlation, there is much to be gained by examining data to see if certain assumptions are tenable. This is particularly true in the case of the multivariate normal model, since it constitutes the basis of most available multivariate statistical methods (see, e.g., Anderson, 1958). Two examples which are of particular importance to psychologists are multiple-correlation methods and the statistical procedures which have been developed in association with factor analysis. With respect to the former it is important to note that if the distribution is not multivariate normal, it follows that either the regressions are non-linear or the variances of the conditional distributions about the regression lines are not homogeneous.[2] As far as factor analysis is concerned, the whole range of maximum likelihood estimation and testing procedures now available (see Lawley and Maxwell, 1963) are based on the multivariate normal model.

---

[1] Now at the Pennsylvania State University.

[2] For a proof of this theorem see Kendall and Stuart (1961, Vol. II, pp. 352–354). The theorem holds unless the variables are mutually completely independent or they are functionally related.

Given the fundamental role of the multinormal model in multivariate statistics it is important to enquire into the feasibility of devising some means of testing the tenability of this assumption for a given set of sample data.

2. DERIVATION OF A POSSIBLE GOODNESS-OF-FIT TEST

The multivariate normal distribution possesses a number of distinctive properties which in principle might form the basis of a goodness-of-fit test for sample distributions. The general difficulty is to find such a property that will lead to a satisfactory test and which will readily generalize beyond two or three variates. Examination of the higher order moments, for example, presents itself as one possibility. Under the null hypothesis of multivariate normality these moments are fixed functions of moments of the first two orders. Apart from the problem of the sampling distribution for these higher order moments, the number of such moments that would need to be considered in any reasonable test rapidly becomes prohibitive as the number of variates increases.

There is, however, one well-known property of the normal distribution which leads to a possible test and which readily generalizes to any number of variates.

Consider the set of p variates represented by the column vector $\mathbf{x}$ or the row vector

$$\mathbf{x}' = [x_1 x_2 \ldots x_i \ldots x_p]$$

and let

$$E(\mathbf{x}') = \boldsymbol{\mu}' = [\mu_1 \mu_2 \ldots \mu_i \ldots \mu_p]$$

be the row vector of their respective means. Let $\mathbf{V}$ be the $p \times p$ variance-covariance matrix for the p variates. If the joint distribution of $\mathbf{x}$ is multivariate normal its probability differential may be written as

$$dF \propto \exp\{-\tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\mathbf{V}^{-1}(\mathbf{x}-\boldsymbol{\mu})\}\Pi\, dx. \tag{1}$$

Then there exists an orthogonal matrix, $\mathbf{H}$, such that the transformation

$$\mathbf{x} = \mathbf{H}\mathbf{y} \tag{2}$$

yields

$$\mathbf{H}'\mathbf{V}^{-1}\mathbf{H} = \mathbf{D},$$

where $\mathbf{D}$ is a diagonal matrix.[3] Since $\mathbf{H}$ is an orthogonal matrix we have from eqn. (2) that

$$\mathbf{y} = \mathbf{H}'\mathbf{x}. \tag{3}$$

---

[3] This is a well-known theorem in matrix algebra (see Kendall and Stuart, 1963, Vol. I, p. 347). I am indebted to Prof. H. O. Lancaster of the University of Sydney for pointing out its relevance in the present context. The theorem is in fact the basis of the method of principal component analysis.

The transformation matrix required to make $\mathbf{D}$ a diagonal matrix is the matrix of eigenvectors of $\mathbf{V}^{-1}$ (which are also the eigenvectors of $\mathbf{V}$). $\mathbf{D}$ is then the diagonal matrix whose elements are the eigenvalues of $\mathbf{V}^{-1}$. It is to be noted that $\mathbf{D}=\mathbf{L}^{-1}$ where $\mathbf{L}$ is the diagonal matrix of the eigenvalues of $\mathbf{V}$ with typical element l_i, say.

Substituting eqn. (2) in eqn. (1) it is easily shown that the latter becomes

$$dF \propto \exp\{-\tfrac{1}{2}(\mathbf{y}-\mathbf{H}'\boldsymbol{\mu})'\mathbf{D}(\mathbf{y}-\mathbf{H}'\boldsymbol{\mu})\}\Pi\, dy.$$

Let $\boldsymbol{\mu}^* = \mathbf{H}'\boldsymbol{\mu}$ be the column vector of the transformed means. Then replacing $\mathbf{D}$ by $\mathbf{L}^{-1}$ we have

$$dF \propto \exp\{-\tfrac{1}{2}(\mathbf{y}-\boldsymbol{\mu}^*)'\mathbf{L}^{-1}(\mathbf{y}-\boldsymbol{\mu}^*)\}\Pi\, dy.$$

Since $\mathbf{L}$ is diagonal this may be rewritten as

$$dF \propto \prod_i \exp\left\{-\tfrac{1}{2}\frac{(y_i-\mu^*{}_i)^2}{l_i}\right\}dy_i. \qquad (4)$$
$$(i=1, 2 \ldots p)$$

Thus under the hypothesis of multivariate normality the eigenvector transformation given by eqn. (2) yields a set of normally and independently distributed variates, y_i. The distribution of the y_i will have means μ_i^* and variances l_i.

To test the hypothesis of multivariate normality one might first examine the original marginal distributions for normality, since if this condition is not met the distribution is clearly not joint normal. Supposing, however, that these marginals are found to be normal or are rescaled to make them normal, the hypothesis of multivariate normality could be tested by replacing $\mathbf{V}$ with the sample covariance matrix, applying the transformation of eqn. (3) and testing the resulting distribution for marginal normality and joint independence.

This latter testing can be accomplished employing the usual multiple contingency chi-square procedures. Under the hypothesis of marginal normality (of the y_i) and complete independence, the total chi-square constitutes a single test of the null hypothesis of multivariate normality.

A difficulty which immediately arises is that unless the number of measures on each variate is very large, many of the expected frequencies for the calculation of the total chi-square will fall below unity as soon as the number of variates becomes at all large. Suppose, following Mann and Wald (1942) and Gumbel (1943), we select marginal class intervals so as to equalize expected frequencies. If there are N such intervals on each marginal, then in order that the expected frequencies in each cell be greater than or equal to a we would require $n=aN^p$ measures on each variate. Even with as few as four class intervals on each marginal and a minimum expected frequency of five, the required sample size rapidly becomes very large. For three variates it is 320, for four variates 1,280. This difficulty is only in part a result of the asymptotic nature of the chi-square distribution as applied to contingency analysis. It is largely a consequence of a difficulty that would confront any goodness-of-fit test; the number of degrees of freedom on which an adequate test must be based rapidly exceeds the bounds of a reasonable sample size. Unless there are many observations constituting points in the multidimensional sample space only the grosser features of the underlying continuous population distribution function can be determined. The problem becomes somewhat analogous to testing the normality of a univariate distribution on the basis of only a few observations.

3. Discussion

Since in psychological data the number of observations is typically small relative to the number of variates, it is clear that in most cases an adequate test of the multivariate normal assumption will be virtually impossible. One is then placed in the highly undesirable (although frequently practised) position of making statistical decisions, the validity of which rest on an untested assumption. In the face of this difficulty two policies are advocated.

(i) *A Partial Testing Procedure* An incomplete test can be accomplished by applying the transformation (3) and then carrying out as much of the contingency analysis as the sample size will permit. In most cases it will be feasible to test only the normality of the transformed marginals, although by collapsing intervals, pair-wise, (and possibly higher order) independence might also be tested. It is suggested that this partial testing will detect most instances of multivariate distributions which have normal marginals but which depart seriously from joint normality. A numerical example of this partial testing procedure is given in the final section of the paper.

(ii) *Studies of Robustness* The second policy advocated is that investigations be conducted into the robustness of relevant statistical methods against violation of the multivariate normal assumption. In cases where an adequate test of a distribution assumption is not possible the issue of robustness becomes one of primary importance. An example of such a study is the recent work of Fuller and Hemmerle (1966) on the robustness of the maximum likelihood estimation procedures in factor analysis.

One of the difficulties associated with studies of robustness is the large number of alternatives to be considered. Violation of the multivariate normal assumption could occur in an indefinite number of ways, and in the multivariate situation it is difficult to specify the kinds of violations most likely to occur in practice. The principles from which the test described in this paper has been derived provide a possible basis for a systematic study of robustness. Violation of the multivariate normal assumption implies that the variates after being transformed according to eqn. (3) are either non-independent or have non-normal marginals. The sensitivity of a given procedure to violation of each of these implications could be assessed.

In particular, variations in the form of the transformed marginal distributions might be systematically studied. Furthermore, each of these distributions is associated with an eigenvalue (its variance) which represents a certain proportion

TABLE 1. BIVARIATE FREQUENCY DISTRIBUTION OF ORIGINAL DATA

(Marginal frequencies of the transformed distribution are given in parentheses)

Mid-pt. of class interval	−2·75	−2·25	−1·75	−1·25	−0·75	−0·25	0·25	0·75	1·25	1·75	2·25	2·75	Marginal frequencies	
2·75								2	3	1			6	(19)
2·25						1	2	6	7		1		17	(30)
1·75					1	1	10	17	11	1	2	1	44	(57)
1·25				1	2	17	27	30	9	3	2	1	92	(63)
0·75			1	2	4	38	50	35	10	5	4	1	150	(91)
0·25			1	3	36	61	36	17	19	13	3	2	191	(155)
−0·25		1	3	6	50	37	40	20	19	11	3	1	191	(258)
−0·75	1	2	7	47	40	15	12	10	8	6	2		150	(202)
−1·25	1	11	26	18	6	8	6	7	5	4			92	(91)
−1·75	4	2	3	12	6	6	5	5	1				44	(28)
−2·25		1	2	2	3	5	3	1					17	(6)
−2·75			1	1	2	2							6	(0)
Marginal frequencies	6 (7)	17 (19)	44 (59)	92 (94)	150 (122)	191 (170)	191 (196)	150 (182)	92 (93)	44 (43)	17 (12)	6 (3)	1000	

of the sum of the variances of the original p distributions. In most psychological data, especially ability measures, there will typically be a considerable discrepancy between the largest and the smallest eigenvalue. It seems reasonable to suppose that the non-normality of a transformed marginal distribution associated with an eigenvalue which accounts for a high proportion of the total variance represents a more serious departure from assumption than does that of a distribution associated with a relatively small eigenvalue.

4. NUMERICAL EXAMPLE

The partial testing procedure based on the method outlined above will be illustrated with a simple bivariate example. This will also serve to show how the procedure may detect a distribution which displays *marginal* normality, yet diverges quite markedly from *joint* normality.

A set of 1,000 contrived scores on each of two variates were standardized and grouped in intervals of 0·5 of a standard deviation. The joint frequency distribution is given in Table 1. The marginals of this distribution conform exactly to the normal curve. Routine calculation gives the following results

$$\mathbf{H} = \begin{matrix} 0{\cdot}7071 & 0{\cdot}7071 \\ 0{\cdot}7071 & 0{\cdot}7071 \end{matrix} \qquad \begin{matrix} l_1 = 1{\cdot}463 \\ l_2 = 0{\cdot}537 \end{matrix}$$

The scores were then transformed according to eqn. (3) and for convenience of presentation were standardized and grouped in the same class intervals as the original distribution. The marginal frequencies of the transformed distribution are given in parentheses in Table 1. The application of a chi-square goodness-of-fit test to these transformed marginal distributions yields the results summarized in Table 2. The hypothesis that these distributions are normal is clearly untenable. Thus the hypothesis that the original distribution is bivariate normal must also be rejected.

TABLE 2. CHI-SQUARE GOODNESS-OF-FIT TEST FOR THE MARGINALS OF THE TRANSFORMED DISTRIBUTION

Transformed Marginal	*Eigen-value*	χ^2	*d.f.*	*P*
1	1·463	28·1	9	$<0{\cdot}001$
2	0·537	155·6	9	$<0{\cdot}001$
Total		183·7	18	$<0{\cdot}001$

REFERENCES

ANDERSON, T. W. (1958). *Introduction to Multivariate Statistical Analysis.* New York: Wiley.

BINDER, A. (1959). Consideration of the place of assumptions in correlational analysis. *Amer. Psychologist*, **14**, 504–510.

CARROLL, J. B. (1961). The nature of the data, or how to choose a correlation coefficient. *Psychometrika*, **26**, 347–372.

FULLER, E. L. and HEMMERLE, W. J. (1966). Robustness of the maximum-likelihood estimation procedure in factor analysis. *Psychometrika*, **31,** 255–266.

GUMBEL, E. J. (1943). On the reliability of the classical χ^2 test. *Ann. math. Statist.* **14,** 253–256.

KENDALL, M. G. and STUART, A. (1963). *The Advanced Theory of Statistics. 2nd Ed.* Vol. I: *Distribution Theory*. London: Griffin.

KENDALL, M. G. and STUART, A. (1961). *The Advanced Theory of Statistics.* Vol. II: *Inference and Relationship*. London: Griffin.

LAWLEY, D. N. and MAXWELL, A. E. (1963). *Factor Analysis as a Statistical Method.* London: Butterworths.

MANN, H. B. and WALD, A. (1942). On the choice of the number of class intervals in the application of the chi-square test. *Ann. math. Statist.* **13,** 306–317.

Vol. 20 Part 1 71–79

The British Journal of Mathematical and Statistical Psychology

May 1967

CLUSTER ANALYSIS OF NON-VERBAL JUDGEMENTS OF FACIAL EXPRESSIONS

By Peter Stringer

University College London

Previous researches on judgements of facial expressions have used either verbal scales, pre-selected or elicited, or non-verbal multidimensional techniques. Certain disadvantages in the methods so far used are described, and a simple alternative procedure is tested. Data from a ' free-grouping ' task are analysed by a clustering method which is congruent with the assumption that the primary response to facial expressions is in terms of contextual or functional similarity. The principal disjoint clusters obtained are described in detail. The meaning of the clusters and of their interrelationships is discussed. It is suggested that there are wider applications both of the ' free-grouping ' procedure and of cluster analysis.

1. Introduction

In recent years the most noticeable body of research on the judgement of facial expressions has been devoted to elucidating the dimensional theory first proposed by Schlosberg (1952, 1954). It has been shown that subjects' responses to emotional expressions can not only be meaningfully interpreted in terms of three dimensions (pleasant-unpleasant, attention-rejection, and sleep-tension), but that subjects can also reliably locate these expressions on scales corresponding to the dimensions (Engen and Levy 1956; Engen, Levy and Schlosberg, 1957, 1958; Triandis and Lambert, 1958).

The weakness of this approach is that by employing only certain pre-selected verbal scales to arrive at dimensions it takes no account of other cognitive tactics that may be importantly employed in extra-experimental situations. In the first place selected scales may include the trivial and irrelevant and exclude the essential, from the viewpoint of an individual's everyday behaviour. The experimenter rather strangely undertakes the onus of anticipating and building into his procedure much of the data which one might expect him to derive from the experiment.

As a partial solution to this problem a list of 300 terms, used spontaneously by subjects in interpreting facial expressions, was subjected to further categorization and selection by judges to arrive at a reduced list of 27 scales (Frijda and Philipszoon, 1963). When these scales were employed, four dimensions were identified, one at least of which (naturalness-submission versus artificiality-condescension) did not correspond to those found by other experiments. This may be because they committed a ' category mistake ', by including in their list scales which probably would be taken to refer to other properties of the face than those relating to temporary emotional state.

Secondly, even if one were to elicit scales from individual subjects, in the manner of the repertory grid technique (Kelly, 1955), the possibility still remains that as verbal scales they may not adequately represent the responses available to a subject in extra-experimental situations; or, alternatively, that they invite the subject to respond in what may be a relatively unusual modality. It can also be objected that the verbal procedure leads to dimensions which explain not so much the underlying structure of facial expressions as of words which are or might be used to denote facial expressions.

Multidimensional scaling has been used to avoid this difficulty (Abelson and Sermat, 1962), with the experimenters asking for similarity ratings on all 78 possible pairs of a set of 13 photographs. Evidence emerged that two of Schlosberg's three scales were sound; the attention-rejection scale proved redundant, and no additional meaningful dimensions were forthcoming. The drawbacks in this investigation lie in the small, though in one sense representative, sample of stimuli used (they were chosen as representative of Schlosberg's original set of 48), and in the nature of the task involved. Even if a small set of stimuli are used, such multidimensional techniques require the subject to perform a task which at best tends to be artificial and tedious; with a large set, subjects need very high motivation indeed to perform.

The present study was prompted by the wish to evolve a simple procedure for the judgement of facial expressions, which was both potentially non-verbal and also involved the subjects in a task which could be carried out quickly and spontaneously for a fairly large set of stimuli. The procedure should be as congruent as possible with everyday, extra-experimental situations; and also with the method of analysis used to derive results.

On the assumption that a facial expression is generally perceived in the context of a series of preceding and succeeding expressions, it was hypothesized that the primary response to one or more expressions would be to construe their contextual (or functional) similarity to other examples (past or present). It would thus be a natural and simple task to require subjects to group a set of facial expressions according to observed similarity. And since cluster analysis offers a non-parametric, descriptive way of grouping functionally similar elements in a matrix of interrelationships, a congruent method of analysis is available for determining the structure of subjects' judgements in this task.

2. EXPERIMENT

Subjects 30 female art-teacher students at Bath Academy of Art, in their first or second year of study. All the subjects were randomly selected; no invitation to participate in the experiment was refused.

Materials The 30 photographs described by Frijda and Philipszoon (1963) were used, and a ' free grouping ' procedure was introduced.

Procedure The complete set of photographs were shown to subjects individually, in a random, open array, and subjects were asked to group these once in such a way that similar facial expressions were included in the same group. They were told that there was no maximum or minimum limit either to the number or size of groups formed. It was emphasized that there was no right or wrong way of doing the grouping; the subject was simply told to group together those photographs which, in her own judgement, were most similar, in such a way that each facial expression was more like those others in its own group than it was like any other one in any other group.

When a subject's results had been tabulated, she was asked whether she was aware of having used verbal labels in the grouping activity; and was required to provide the most appropriate label for each group formed.

The entire judgement session lasted approximately 20 minutes.

Cluster Analysis of Data A variant (Cluster 3) of a general purpose clustering program described by Constantinescu (1966, 1967) was used. The method uses graphs and trees and their matrix representations, and describes the interrelations of a set of elements which may have a number of functionally common properties.

In the present instance the inferred similarity frequencies for all possible pairs of the 30 stimuli were calculated. A matrix was derived by placing a count of 1 in the appropriate cell on each occasion that a subject grouped any one photograph with any other photograph. Cell frequencies could thus vary from 0 to 30, and were taken as the overall measure of judged similarity of all pairs of facial expressions.

The *distances* between elements (that is, the frequencies with which pairs of elements were *not* associated together in the same group) were entered in the matrix M (Table 1). 'Partially disjoint' clusters of stimuli as grouped by subjects were derived from M by means of Cluster 3 (Table 2). The final pattern in Table 2 was produced by an Atlas computer in approximately five minutes.

3. RESULTS

Subjects produced variously from 4 to 13 groups (mean = 8·40), with from 1 to 12 photographs in each group (mean = 3·57). Almost all subjects reported that they found the task 'easy' and 'natural', although some felt that they could have elaborated and qualified their judgements given considerably more time. Just over half reported the conscious use of verbal labels while forming groups.

The search for principal disjoint clusters (not only on a given level) revealed five: (A) Worry—(stimuli numbers) 6, 8, 9, 12 (level 18); (B) Disgust, Pain—7, 11, 17, 20, 28 (level 24); (C) Surprise—2, 3, 4, 6, 17, 18, 19, 22 (level 25); (D) Thoughtfulness—10, 14, 21, 23, 24, 26, 29; (E) Happiness—1, 5, 13, 15, 16, 25, 27, 30 (level 28). The level number indicates the level at which all the elements were absorbed in the cluster, and corresponds to the value of the largest inter-element distance in the cluster. It will be noted that clusters A and C and B and C are not strictly disjoint; also cluster D does not in fact occur at any level of the cluster solution. These points will be amplified below. The verbal

TABLE 1. MATRIX OF DISTANCES BETWEEN ELEMENTS

STIMULUS NUMBER (ELEMENTS)

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**
1	—	30	29	30	23	30	30	30	29	26	29	29	12	25	24	17	30	28	29	30	28	30	29	26	12	27	12	30	29	25
2		—	4	10	30	23	28	29	28	29	26	28	29	27	27	30	22	13	21	26	30	15	30	29	29	30	30	29	28	29
3			—	9	29	23	29	29	28	28	27	29	29	29	29	30	20	16	17	28	30	15	30	30	30	30	30	29	30	29
4				—	28	21	29	27	27	29	28	28	29	28	29	30	21	16	19	29	28	16	29	29	29	29	30	28	29	29
5					—	29	29	29	29	30	29	29	18	30	28	17	29	29	29	29	29	29	30	30	27	30	20	30	29	11
6						—	26	16	14	29	27	17	30	28	30	30	23	22	19	25	30	25	30	29	30	28	30	25	28	29
7							—	27	27	30	10	28	30	30	29	30	14	29	27	7	30	22	29	30	30	30	30	16	29	27
8								—	15	26	28	18	30	28	28	29	27	28	27	25	27	29	27	26	29	25	30	21	22	30
9									—	30	29	17	29	29	29	30	28	24	25	28	29	28	28	28	30	26	29	22	27	30
10										—	30	21	29	19	26	29	30	29	26	29	11	30	19	19	26	20	30	30	26	30
11											—	29	30	30	27	30	18	28	27	8	30	21	30	30	30	30	30	17	29	24
12												—	29	28	29	30	28	28	23	28	23	28	25	27	30	26	29	28	26	30
13													—	28	23	11	29	28	29	30	30	29	28	29	18	29	7	30	28	19
14														—	28	30	30	27	30	30	24	30	21	12	21	18	29	29	20	30
15															—	20	29	27	28	29	26	29	20	27	23	28	21	30	26	28
16																—	30	30	30	30	29	30	26	30	20	28	10	30	29	18
17																	—	25	22	17	30	10	30	30	30	30	30	22	30	28
18																		—	22	28	28	19	29	28	28	29	29	29	28	30
19																			—	28	26	22	29	30	30	29	30	28	30	29
20																				—	30	23	30	30	30	29	30	17	28	25
21																					—	30	16	20	28	18	30	30	26	30
22																						—	30	30	30	30	30	26	30	29
23																							—	18	25	26	27	29	21	30
24																								—	20	13	29	28	16	30
25																									—	23	16	30	26	28
26																										—	27	29	18	30
27																											—	30	28	21
28																												—	25	27
29																													—	30
30																														—

TABLE 2. FINAL PATTERN

The final pattern includes only those subsets (clusters) on each level which are ' partially disjoint ' (any initial element in a subset is not in any precedent subset on the same level) (Constantinescu 1966). The level number corresponds to the value of the largest inter-element distance in the cluster.

LEVEL	*CLUSTERS OF ELEMENTS*				
4	2, 3				
7	7, 20	13, 27			
8	11, 20				
9	3, 4				
10	2, 3, 4	7, 11, 20	16, 27	17, 22	
11	5, 30	10, 21	13, 16, 27		
12	1, 13, 27	14, 24			
13	24, 26				
14	6, 9				
15	2, 3, 22	8, 9			
16	2, 3, 4, 22	6, 8, 9	21, 23	24, 29	25, 27
17	1, 13, 16, 27	17, 20	20, 28	23, 26	
18	5, 13, 16	6, 8, 9, 12	7, 11, 17, 20	14, 24, 26	21, 23, 26
19	2, 3, 4, 18, 22	5, 13, 16, 30	10, 14, 24	26, 29	
20	1, 13, 16, 25, 27	14, 24, 26, 29			
21	3, 4, 17, 22	5, 13, 16, 27, 30	21, 23, 24, 26		
22	2, 3, 4, 17, 19, 22	23, 24, 26, 29			
23	1, 5, 13, 16, 27	2, 3, 4, 6, 17, 19	24, 25, 26		
24	7, 11, 17, 20, 28				
25	1, 5, 13, 16, 27, 30	2, 3, 4, 6, 17, 18, 19, 22	23, 24, 25, 26	28, 29	
26	6, 7, 17, 20, 22, 28	21, 23, 24, 26, 29			
27	1, 5, 13, 16, 25, 27	4, 6, 8, 9, 19			
28	1, 5, 13, 15, 16, 25, 27, 30	17, 18, 19, 20, 22			
29	1, 3, 5, 9, 11, 12, 15, 18, 19	16, 21, 23, 25, 26, 29			
30	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30				

labels were arrived at by considering all the labels attached by subjects to any groups containing at least one photograph from a particular cluster. In other words, the 30 labels given to each photograph as a member of a group were 'summed' to produce first an 'average' label for that element (Table 3); then these labels were 'summed' for each cluster to produce a cluster-title. There was considerable general agreement among subjects over the labels; most of those given to any single photograph fell into a small number (one, two or three) of clearly definable and usually related categories.

TABLE 3. VERBAL LABELS GIVEN TO FACIAL EXPRESSION STIMULI

This table lists the main categories of verbal labels given to each stimulus, and is only a summary of actual labels used in many cases. Nuances referred to in the text are omitted for the sake of economy.

STIMULUS	VERBAL LABELS	STIMULUS	VERBAL LABELS
1	happy, pleased; calm	16	happy, pleased, amused
2	surprise	17	surprise; horror, pain; fear
3	surprise	18	surprise
4	surprise	19	surprise; anger, rebellion
5	amused, happy	20	disgust; horror, pain; misery
6	doubt, worry	21	pensive; stubborn, aloof; angry
7	disgust; pain	22	surprise; horror; fear
8	worry; sadness; questioning	23	thoughtful; disdainful
9	worry; doubt	24	thoughtful, calm
10	pensive; stubborn; questioning	25	happy; thoughtful
11	disgust	26	thoughtful; stubborn, aloof
12	worry; thoughtfulness; anger	27	happy, pleased
13	happy, pleased	28	pain; sorrow; aversion
14	pensive, dreamy	29	thoughtful
15	sceptical; thoughtful; cheeky	30	happy, laughing.

The nucleus of cluster A is made up of elements 6, 8, 9 which represent a puzzled, questioning, and sometimes surprised, worry. 12 seems to be a more concentrated expression of worry, verging towards anger.

Cluster B has elements 7, 11, 20 in its main sub-cluster, which chiefly represents disgust (distaste, aversion, etc.) and pain. 17 and 28 have more of the 'pain' component, 'horror-pain' and 'sad-pain' respectively. Since 20 has more equal parts of disgust and pain with some misery it serves to link 7 and 11 with 17 and 28—the axis of an 'aversion' cluster running from distaste to various types of pain.

In cluster C, 6 belongs rather tenuously, as can be seen by examining its similarity scores with other elements of the cluster (Table 1). One may note that it would not have been included in this cluster had 17 and 18 been more similar (for in that case all the other elements would have been absorbed before 6). It may be preferable to discount it from this cluster, and treat A and C as disjoint. Elements 2, 3, 4 are the 'pure' component of the cluster, representing surprise (astonishment, shock). 17 and 19 represent respectively horror and fear, and anger, in addition to surprise. 22 is closely related to both 17 and 19 (despite the lack of an ‘ anger ’ component among its labels), as well as to 2, 3, 4, and this acts as the axis of this cluster. Element 18 is more particularly akin to 2, 3, 4 than to the others. A cluster of 2, 3, 4, 18, 22 (level 19) is an intermediary cluster between the nuclear 2, 3, 4, (level 10) and the whole cluster. Cluster C runs from ‘ pure ’ surprise to surprise tinged with horror and fear, or anger.

Cluster D is slightly more complex. The nuclear cluster comprises elements 21, 23, 24, 26, 29 (occurring in level 26), which resolves into two higher-level clusters—21, 23, 26 (level 18) and 24, 26, 29 (with 14, level 20). The former seems to represent a positive thoughtfulness (overtones of disdain, defiance, anger) mixed with determination, while the latter is described as a calmer, more contemplative thoughtfulness, with a suggestion of worry and sorrow (particularly 29). Element 26 sufficiently combines both aspects. 14 occurs in a cluster with 24, 26, 29, but is not linked with 21 and 23. It represents a particular example of peaceful, dreaming thoughtfulness. 10 combines determination, stubbornness, and questioning, with thoughtfulness; and is accordingly most similar to element 21. Although in the present solution it is not absorbed into any of the lower-level clusters, as 14 is, examination of the similarity matrix will show that it fairly belongs to cluster D. While 10 and 21 represent one side of this cluster, 14 and 24 represent the other.

Happiness, ranging from calm pleasure to hilarious amusement, is indicated by the labels given to elements in cluster E. The former is represented particularly by 1 and 25, the latter by 5 and 30. The higher-level cluster 13, 16, 27 (level 11) combines both components to form a link between these two pairs; it verges rather in the direction of the former, particularly in respect of the labels attributed. Element 15 (scepticism, thoughtfulness, cheekiness) is uneasily integrated with this group, although it seems to combine thoughtfulness and joking; in this context it clearly is not of the same order as other elements in clusters D and E.

Where there was a similarity frequency of 7 or more for any extra-cluster pair, that is between elements of different clusters, this was taken as an indication of inter-cluster relationships. The following were found: 8 and 9 with 28 (clusters A and B); 12 with 19 (A and C); 8 with 29, 12 with 10 and 21 (A and D); 7, 11, 20 with 22 (B and C); 14, 24, 26 with 25, 23 with 15 (D and E). (The presence of elements 6 and 17 in clusters A and C and B and C respectively has been mentioned above.) Again from the evidence of the labels, it is possible to infer a description of these relationships. ‘ Disgust ’ and ‘ surprise ’ are linked by a component of ‘ horror ’, and are respectively related to ‘ worry ’ by way of ‘ pain ’ and ‘ sadness ’, and by an ‘ angry questioning ’. ‘ Worry ’ is joined to ‘ thoughtfulness ’ in respect of both a ‘ questioning worry ’ and an ‘ angry concentration ’ or ‘ stubbornness ’ (it should be remembered that there are ‘ anger ’ components in cluster A). ‘ Thoughtfulness ’ and ‘ happiness ’ are joined simply by shared elements of thoughtfulness and calmness.

4. DISCUSSION

These results cannot strictly confirm or weaken those of other studies, which have generally used factor methods of analysis. Cluster analysis identifies clustered patterns of functionally similar variables, rather than measuring component influences. A single cluster may contain several components, and a single component may be represented in several clusters. Furthermore, the description of a cluster's meaning in the present experiment must differ in kind from the interpretations given to factors; the former is educed from the way in which subjects labelled the groups they formed, while the latter is a logical consequence of the subject's use of pre-determined verbal scales.

However, in view of the remarks already made on the structure of elements in the present experiment, it is possible to describe an approximate line running from hilarity and happiness through thoughtfulness, worry and anger, to surprise and fear, *or* pain and disgust, and ultimately perhaps to horror. This line clearly reflects the ' pleasant-unpleasant ' dimension of previous experiments, as well as the opposition of control to intensity of expression. Controlled expressions (thoughtfulness) seem to occupy a rather more axial than polar position in the overall pattern. ' Sleep-tension ' and ' attention-rejection ' (or disinterest) are not exactly represented. Instead there is an opposition between active and passive expressions, between the axial thoughtfulness, worry and anger, and other clusters. (Active expressions reflect thoughts or emotions directed at, rather than produced by, a feature of the inner or outer environment.) Frijda and Philipszoon's (1963) unique factor of ' naturalness and submission versus artificiality and condescension ' is perhaps present here as an elaboration of ' passive versus active '; for ' active ' elements labelled ' thoughtfulness ', ' worry ' and ' anger ', and only these elements, were also labelled ' aloof ', ' disdainful ', and sometimes were considered to be ' posed ', ' social ', or ' phoney '.

5. CONCLUSION

It is possible, by the experimental procedure tested here, to obtain a detailed and intelligible description of the way subjects structure a set of facial expressions. Five main clusters with their internal and external linkages have been described. How stable these clusters are, and how many additional ones might be found in a larger set of stimuli, can only be determined by further experiments.

Studies are in progress to test the assumption that the primary response to facial expressions is in terms of their contextual similarity; and to compare the present cluster analysis with solutions derived by other methods from the same data.

Since one of the criteria of Frijda and Philipszoon (1963) in choosing their 30 stimuli was representativeness, we may ask how far present results confirm that their criterion was met. In so far as there is imbalance in the size of clusters and in the triangular structure of clusters A, B and C, it was not met.

Structural anomalies within clusters may also rule against the claim. The method used here might be extended to the specific study of stimulus equivalence and representativeness.

It would be possible to study subject typology in respect of responses to the " free grouping " task, by determining the frequencies with which any pair of subjects produced the same inferred judgement of similarity. A matrix could be formed and cluster analysed. The number and size of groups they produced, as well as their differing ease in arriving at verbal labels, might be interesting variables. The labels themselves could be used in a semantic differential situation, judging the same stimuli, for methodological comparison.

For the future this study suggests the investigation of facial expression judgements by more varied techniques, particularly simple, non-verbal techniques; the extension of the ' inferred similarity ' method to different material and experimental situations; and a wider application of cluster analysis as a descriptive instrument.

ACKNOWLEDGEMENTS

I am indebted to Dr. Paul Constantinescu, of the University of Bucharest, for writing the clustering program used in this study; and for his advice and encouragement in many discussions of cluster analysis.

REFERENCES

ABELSON, R. P. and SERMAT, V. (1962). Multidimensional scaling of facial expressions. *J. exp. Psychol.* **63,** 546–554.

CONSTANTINESCU, P. (1966). The classification of a set of elements with respect to a set of properties. *Computer Journal*, **8,** 352–357.

CONSTANTINESCU, P. (1967). A method of cluster analysis. *Br. J. math. statist. Psychol.* **20,** 93–106

ENGEN, T. and LEVY, N. (1956). Constant sum judgments of facial expressions. *J. exp. Psychol.* **51,** 396–398.

ENGEN, T., LEVY, N. and SCHLOSBERG, H. (1957). A new series of facial expressions. *Amer. Psychologist*, **12,** 264–266.

ENGEN, T., LEVY, N. and SCHLOSBERG, H. (1958). The dimensional analysis of a new series of facial expressions. *J. exp. Psychol.* **55,** 454–458.

FRIJDA, N. H. and PHILIPSZOON, E. (1963). Dimensions of recognition of expression. *J. abnorm. soc. Psychol.* **66,** 45–51.

KELLY, G. A. (1955). *The Psychology of Personal Constructs.* New York: W. W. Norton and Co. Inc.

SCHLOSBERG, H. (1952). The description of facial expressions in terms of two dimensions. *J. exp. Psychol.* **44,** 229–237.

SCHLOSBERG, H. (1954). Three dimensions of emotion. *Psychol. Rev.* **61,** 81–88.

TRIANDIS, H. C. and LAMBERT, W. W. (1958). A restatement and test of Schlosberg's theory of emotion with two kinds of subjects from Greece. *J. abnorm. soc. Psychol.* **56,** 321–328.

Vol. 20 Part 1 81–92
The British Journal of Mathematical and Statistical Psychology
May 1967

A METHOD FOR INVESTIGATING CHILDREN'S UNDERSTANDING OF CERTAIN LOGICAL CONNECTIVES USED IN BINARY PROPOSITIONAL THINKING

By E. A. Peel
University of Birmingham

A theoretical frame of reference and an experimental method are described for investigating children's understanding of such logical connectives as *If . . . then* in certain binary propositions. A game is played with coloured beads and counters between experimenter and child which both obviates the need to use 'naturally' connected binary propositions and at the same time prevents the rejection of the propositions on account of their arbitrariness. The games consist essentially of a bead draw by the experimenter followed by a responding draw of a counter by the child, according to the instructions for each game.

The frequencies of different conjunctions of bead and counter draws (pq, $p\bar{q}$, $\bar{p}q$, $\overline{pq}$, where p and q are different kinds of beads and counters respectively) are obtained and converted to proportions. These proportions are compared with those expected theoretically according to the definitions of the main binary propositions.

1. Introduction

There have been few previous studies of what young children understand by: *If . . . then*, *If . . . then not*, and *Either . . . or* (Matalon, 1962; Naess, 1962; Hill, 1961; and Suppes, 1965), although there have been more extensive studies of the syllogistic reasoning of children and adults (Mays, 1963; Hill, 1961; Suppes, 1965; and Wilkins, 1929).

When investigating what young children understand by the various connectives, a real difficulty is found in having to model the propositions in such a way as to minimize the dependence of the method on language and meaningful associations. Also, in the case of the first set of connectives, *If . . . then*, there is the well known tendency, revealed by most people, to require a meaningful relation between the two propositions connected by *If . . . then*. Thus they will accept such binary propositions as *If Kate switches on the television, then her brother goes out*, but find it more difficult to accept *If cats have claws, then the universe is expanding*. It seems that the logician's purely formal implication has usually to be encased in a natural implication in which the second proposition can be seen to depend naturally or causally upon the first. This has led both logicians and psychologists to distinguish between these two ideas of implication. Matalon (1962) investigates them separately by first using an arbitrary situation, involving the switching on of two different coloured lights for what he calls 'material implication'—that is purely formal implication—and secondly by using the natural setting of a village, in which the post office can only be reached by passing a certain house, for investigating what he calls 'natural implication'. However, it is possible to bridge the gap and to produce arbitrary propositions which do not outrage the sense of reality of the thinker, but which are none the less not dependent for their acceptance upon meaningful associations between the constituent propositions. This bridge can be effected by a games technique. The essence of an organized game is that arbitrary rules and relations are accepted, as in the rule that the ace takes the queen, or that if the ball is knocked forward in a game of rugby football a scrum must be had. Such a technique has quite definite advantages when investigating young children, as they are quite prepared to accept arbitrary rules in a situation which appears to have the qualities of a game.

There is also a further difficulty when such binary relations are investigated by ordinary propositions, as in the case of the first one mentioned above:

> *If Kate switches on the television, then her brother goes out.*
>
> *Kate switches on the television.*
>
> *Does her brother go out?*

and that is that one has to take great care that a conclusion is not reached merely because it looks plausible. Thus in her study of the syllogistic reasoning of students, Wilkins (1929) showed that if a conclusion is naturally plausible, it is often accepted even though it does not follow from the premisses given. Similarly, Suppes (1965) draws attention to the same difficulty in the work conducted by Hill (1961) where a badly constructed reasoning problem is one in which the conclusion appears naturally plausible, such as *Boys run faster than girls*.

Formal implication, *If . . . then*, can be dressed out as follows: (*a*) by two naturally and causally related propositions; (*b*) by two unconnected propositions, so put together as to bring about a rejection of the implication, such as *If cats have wings, then* $2+2=4$; and (*c*) by two arbitrarily connected propositions, *If you pass the ball forward, then your side will be penalized*, which are accepted by the thinker as valid rules for a game.

This paper is concerned with a method which utilizes the third type, and so avoids the rejections aroused in (*b*) and is free of the assumptions of natural causes implied in (*a*). Furthermore, by not calling for judgements of written or spoken propositions, it avoids a situation where a conclusion may be affirmed or denied because it is plausible or otherwise. By the game technique, it is hoped to make the binary relations acceptable to the child, and at the same time to enable us to concentrate on a binary relation independently of any natural relation between the two propositions involved. This approach, of course, does not deny that a child's understanding of *If . . . then* has been acquired as a result of natural usage within the family and with other children. Children are often subjected to popular meanings of implication, such as *If you eat your pudding,* *I'll give you a sweet*, which do not at all carry the meaning assumed by the logician. It is the precise purpose of this method to investigate what the child understands by *If . . . then*, *If . . . then not*, and *Either . . . or*, but it is better to do this in a novel situation not tied by previously learned, acceptable relations, and not dependent, as are the methods of Hill (1961), Suppes (1965) and Wilkins (1929), upon an understanding of complex propositions in the form of ordinary English sentences.

2. THE GAME PROCEDURE

The three kinds of binary proposition, *implication*, *incompatibility* and *disjunction*, were modelled in a game played between the experimenter and each individual child. The experimenter has before him a collection of many beads in a tin, there being at least 10 each of the four colours, blue, yellow, red and green. In another shallow tin is a similar collection of counters, also made up of some 10 each of the same colours. The experimenter and the subject face each other with their beads and counters before each respectively, and an empty box is placed on the table between them into which they make their play. The game consists essentially of the experimenter drawing a certain bead, and then the subject drawing a counter depending upon the rule of the particular game. The rules for each of the three games are briefly as follows:

1. Implication: " We may each pick any colour we like, but in this game, if and whenever I draw a red bead, you also draw a red counter."
2. Incompatibility: " We may each draw any colour we like, but in this game, if and whenever I draw a red bead, you are not to draw a red counter."
3. Disjunction: " We both draw so that there is at least one red in the box between us. I shall draw first, and put my bead in, and then you draw a counter and put that in."

The box is emptied after each draw in each game and the counters and beads put back in their respective boxes. Ten pairs of draws are made with each child in a set random order consisting of six reds and four non-reds. The 10-bead draws are initial draws made by the experimenter, to which the child has to make the response by drawing an appropriate counter. The child's draws are recorded in relation to the initial draws made by the experimenter and so a complete record of all possible conjunctions of bead and counter draws is kept.

The game is also played in the *reverse* form in which the child is asked to make an initial bead draw and the experimenter then makes a consequent counter draw. The child is then asked to say whether the experimenter's draw is right or wrong. By this technique the investigator is able to concentrate on the rarer occurring conjunctions made by the child in the binary games of implication, incompatibility and disjunction.

3. Psychological Requirements Relating to Motivation and Reinforcement

First, as to motivation, the testing was conducted in an informal atmosphere. The games aspect was heavily stressed in order to keep up the subject's interest and enthusiasm. Children of the ages of 5 to 10 took a great interest in the competitive aspects of the game, that is, in the competition between each child and the experimenter. It was found that very little needed to be said to maintain this interest for such a short run as 10 draws.

Such reinforcement as was given was not to obtain certain responses, but merely to ensure an act of responding (up to 10 draws). It was therefore important that the reinforcement should be general, pleasant and neutrally uninformative regarding the draw made by the child. It was found quite easy to do this by a pleasant nod of approval, or an utterance of " Yes " or " Fine " and so on. Such a reinforcing situation would not have endured over a long run of draws, but it was found that 10 draws by no means strained the child when no information was given about the correctness or incorrectness of his response. The purpose of the 10 trials was to obtain a more complete picture of the child's responses in order to obtain as many as possible of the binary conjunctions involved, so that an adequate analysis of the group of children as a whole could be undertaken.

Some 150 infant and primary school children played the games and they ranged in age class intervals from 5+ to 11+ with frequencies in each year class ranging from 13 to 30. Since each child responded by making 10 draws, there were, in all, some 1,500 conjunctions to analyse.

4. Theoretical Basis of the Games

The operative colour used in the draws was red. Let us write R for a red bead draw and $\bar{R}$ for a non-red bead draw, made by the experimenter. Let us also write r for a red counter draw made in response by the child and $\bar{r}$ for a non-red counter draw. Non-red means blue, yellow or green. There are, then, four possible conjunctions of bead and counter draws, Rr, $R\bar{r}$, $\bar{R}r$ and $\bar{R}\bar{r}$, each representing a game between experimenter and child. Clearly, the total frequency of occurrence of such conjunctions for any group of children can be counted. Thus for the group of fifteen 5 to 6 year-olds, each of whom played 10 games, the total frequency would be $f_{Rr}+f_{R\bar{r}}+f_{\bar{R}r}+f_{\overline{Rr}}=150$. Furthermore, we may divide these frequencies up according to whether they are based on initial R and $\bar{R}$ draws as follows. There would be 90 initial R draws and 60 $\bar{R}$ draws by the experimenter. Hence $f_{Rr}+f_{R\bar{r}}=90$ and $f_{\bar{R}r}+f_{\overline{Rr}}=60$. We may reduce these frequencies to proportions as follows: $p_{Rr}=f_{Rr}/90$ and $p_{R\bar{r}}=f_{R\bar{r}}/90$ and so on.

The three binary propositions of implication, incompatibility and disjunction will be considered in terms of these proportions. The three propositions will be defined in the manner of the logician, namely:

Implication		$R \supset r$	
	by	$Rr \vee \bar{R}r \vee \bar{R}\bar{r}$,	$R\bar{r}$ being false;
Incompatibility		$R \supset \bar{r}$	
	by	$R\bar{r} \vee \bar{R}r \vee \bar{R}\bar{r}$,	Rr being false;
Disjunction		$R \vee r$	
	by	$Rr \vee R\bar{r} \vee \bar{R}r$,	$\bar{R}\bar{r}$ being false;

and use these as hypotheses to compare with the incidence of binary conjunctions, Rr, $R\bar{r}$, $\bar{R}r$, $\bar{R}\bar{r}$, *actually* arising in the course of the games.

Taking first *implication*, *If . . . then*, as logically defined, it would be expected that all the children's counter draws to R bead draws would be r, with *no* $\bar{r}$ draws ($R\bar{r}$ being false) and that there would be an equal distribution of r and $\bar{r}$ counter draws to initial R draws.

The above binary conjunctions, in terms of the *forward* games, are shown in Table 1:

TABLE 1. IMPLICATION: PROPORTIONS OF CONJUNCTIONS

Initial bead draw by experimenter	R		$\bar{R}$	
Responding counter draw by child	r	$\bar{r}$	r	$\bar{r}$
Binary conjunction	Rr	$R\bar{r}$	$\bar{R}r$	$\bar{R}\bar{r}$
Truth value according to the logical meaning of implication	true	false	true	true
Theoretically expected proportion of each conjunction according to the above truth values	1	zero	0·5	0·5

Similarly, the theoretically expected proportions for the remaining games of incompatibility and disjunction may be obtained and are shown in Table 2.

It is worth-while also to set out the theoretical proportions expected for the binary propositions of logical *equivalence* and *mutual exclusion*, since these two prove to be very significant in the results obtained (Table 3).

In essence, the analysis of the data takes the form of obtaining the observed proportion of occurrence of the four conjunctions, Rr, $R\bar{r}$, $\bar{R}r$ and $\bar{R}\bar{r}$ and comparing these with the above theoretical proportions.

The theoretical proportions of the Rr, $R\bar{r}$, $\bar{R}r$ and $\bar{R}\bar{r}$ conjunctions may also be computed for the *reverse* game. In this game there is an initial bead draw by the child, a responding counter draw by the experimenter and a final acceptance by the child of this counter draw as being correct or incorrect. In the reverse game the number of acceptances and non-acceptances by the child are counted and the proportions are based on these data.

TABLE 2. INCOMPATIBILITY AND DISJUNCTION: PROPORTIONS OF CONJUNCTIONS

Initial bead draw by experimenter	R		$\bar{R}$	
Experimenter/child binary conjunction	Rr	$R\bar{r}$	$\bar{R}r$	$\bar{R}\bar{r}$
INCOMPATIBILITY	false	true	true	true
Theoretically expected proportions	zero	1	0·5	0·5
DISJUNCTION	true	true	true	false
Theoretically expected proportions	0·5	0·5	1	zero

TABLE 3. EQUIVALENCE AND MUTUAL EXCLUSION: PROPORTIONS OF CONJUNCTIONS

Initial bead draw by experimenter	R		$\bar{R}$	
Experimenter/child binary conjunction	Rr	$R\bar{r}$	$\bar{R}r$	$\bar{R}\bar{r}$
EQUIVALENCE ($Rr \vee \bar{R}\bar{r}$)	true	false	false	true
Theoretically expected proportions	1	zero	zero	1
MUTUAL EXCLUSION ($R\bar{r} \vee \bar{R}r$)	false	true	true	false
Theoretically expected proportions	zero	1	1	zero

The procedure and the expected proportions of acceptances and non-acceptances can be tabulated according to the five binary propositions required to analyse the results of the forward game. In this case the proportions are based in each case on the number of instances of each binary conjunction taken separately (Table 4).

Lastly, the matching of non-reds in the forward game is considered, that is the tendency, when the experimenter draws a blue, for the child also to draw a blue, and so on, in which case the non-red colour chosen by the experimenter is assumed to exercise an influence on the counter chosen by the child. This matching turned out to be quite significant in the results and it is therefore necessary to calculate the proportions expected by chance. A count may be made of the total frequencies of matched $\bar{R}\bar{r}$ draws (blue-blue, yellow-yellow, green-green) f_m and of non-matched $\bar{R}\bar{r}$ draws (blue-green, yellow-green, blue-yellow, etc.) $f_{\bar{m}}$ for each of the three logical games. These frequencies can be converted to the proportions:

$$\frac{f_m}{f_m+f_{\bar{m}}}=p_m \quad \text{and} \quad \frac{f_{\bar{m}}}{f_m+f_{\bar{m}}}=p_{\bar{m}}.$$

It is evident that by chance, with three non-red colours, it is expected that $p_m = 0{\cdot}33$ and $p_{\bar{m}} = 0{\cdot}67$ for all three logical games. These proportions are used as a base line to examine the tendency for children to match non-reds.

TABLE 4. THEORETICAL PROPORTIONS OF THE ACCEPTABILITY AND NON-ACCEPTABILITY OF THE CONJUNCTIONS

Initial bead draw by child	R				$\bar{R}$			
Conjunction of the above bead draw with the responding counter draw made by the experimenter	Rr		$R\bar{r}$		$\bar{R}r$		$\bar{R}\bar{r}$	
Decision of the child about the experimenter's draw of a counter	acc.	non-acc.	acc.	non. acc.	acc.	non-acc.	acc.	non-acc.
Theoretically expected proportions								
Implication	1	zero	zero	1	1	zero	1	zero
Incompatibility	zero	1	1	zero	1	zero	1	zero
Disjunction	1	zero	1	zero	1	zero	zero	1
Equivalence	1	zero	zero	1	zero	1	1	zero
Mutual exclusion	zero	1	1	zero	1	zero	1	zero

5. ILLUSTRATION OF THE METHOD ON SOME RESULTS

In order to illustrate the working of the method, some results for children of different chronological age groups are now presented.

Table 5 shows the frequencies of the bead-counter conjunctions in the forward game of implication.

TABLE 5. IMPLICATION GAME: FREQUENCIES OF BEAD-COUNTER CONJUNCTIONS

Initial bead draw			R		$\bar{R}$	
Bead-counter conjunctions		Sample Size	Rr	$R\bar{r}$	$\bar{R}r$	$\bar{R}\bar{r}$
Age group	5+	15	83	6	3	57
	8+	13	77	1	3	49
	11+	30	179	1	4	116

These frequencies give rise to the proportions shown in Table 6.

From Table 6 it seems that the children reacted to the game instruction *If . . . then* as if it meant *If and only when . . . then*. When they were in this situation they were called upon to produce the consequence to the antecedent choice made by the experimenter (see Discussion).

TABLE 6. IMPLICATION GAME: PROPORTIONS OF BEAD-COUNTER CONJUNCTIONS

Bead-counter conjunctions		Rr	$R\bar{r}$	$\bar{R}r$	$\bar{R}\bar{r}$
Theoretically expected proportions based on					
(a) *Implication*		1	zero	0·5	0·5
(b) *Equivalence*		1	zero	zero	1
Age group	5+	0·92	0·07	0·05	0·95
	8+	0·99	0·01	0·02	0·98
	11+	0·99	0·01	0·03	0·97

For the reverse game of implication, played with the same children, the proportions of accepted and non-accepted bead-counter conjunctions shown in Table 7 were obtained.

TABLE 7. PROPORTIONS OF ACCEPTED AND NON-ACCEPTED BEAD-COUNTER CONJUNCTIONS IN THE REVERSE GAME OF IMPLICATION

Initial bead draw		R				$\bar{R}$			
Bead-counter conjunctions		Rr		$R\bar{r}$		$\bar{R}r$		$\bar{R}\bar{r}$	
Judgement of child		acc.	non-acc.	acc.	non-acc.	acc.	non-acc.	acc.	non-acc.
Theoretically expected proportions									
Implication		1	zero	zero	1	1	zero	1	zero
Equivalence		1	zero	zero	1	zero	1	1	zero
Age group	5+	1·00	zero	0·03	0·97	0·24	0·76	0·93	0·07
	8+	1·00	zero	0·01	0·99	0·58	0·42	1·00	zero
	11+	1·00	zero	zero	1·00	0·82	0·18	0·98	0·02

Apparently, here we see an effect, increasing by age, of the 'repetition' by the experimenter of r counter draws to the child's initial $\bar{R}$ bead draws. The conjunction $\bar{R}r$ has a key role in transforming an *If and only when . . . then* binary proposition (equivalence) to one of *If . . . then* (implication).

In order to illustrate the game of incompatibility, the results of the reverse game are shown in Table 8.

Here it is seen that even the young child accepts incompatibility as defined, and this finding is also clearly shown in the forward game. Apparently, although he has difficulty in accepting $\bar{R}r$ in the *If . . then* situation, he has no difficulty at all in accepting the analogous $\bar{R}\bar{r}$ in the *If . . . then not* problem.

Lastly, the results of the method in the forward disjunction game are shown in Table 9.

Here it can be seen how difficult it was for the child to cope with the *at least one red* instruction. At 5+ the best logical approximation is that of the

TABLE 8. REVERSE GAME OF INCOMPATIBILITY

Initial bead draw		R				$\bar{R}$		
Bead-counter conjunctions		Rr		$R\bar{r}$		$\bar{R}r$		$\bar{R}\bar{r}$	
Judgement of child		acc.	non-acc.	acc.	non-acc.	acc.	non-acc.	acc.	non-acc.
Theoretically expected proportions for logical incompatibility		zero	1	1	zero	1	zero	1	zero
Age group	5+	0·34	0·66	0·84	0·16	0·76	0·24	0·87	0·13
	8+	0·07	0·93	1·00	zero	0·86	0·14	0·96	0·04
	11+	zero	1·00	1·00	zero	0·96	0·04	0·96	0·04

TABLE 9. FORWARD GAME OF DISJUNCTION

Initial bead draw		R		$\bar{R}$	
Bead-counter draw conjunctions		Rr	$R\bar{r}$	$\bar{R}r$	$\bar{R}\bar{r}$
Theoretically expected proportions according to logical disjunction		0·5	0·5	1	zero
Theoretically expected proportions according to logical tautology ≡ lack of any system of counter responding		0·5	0·5	0·5	0·5
Theoretically expected proportions according to logical mutual exclusion		zero	1	1	zero
Age gruop	5+	0·60	0·40	0·09	0·91
	8+	0·54	0·46	0·53	0·46
	11+	0·08	0·92	0·89	0·11

disjunction, not of R and r, but of R and $\bar{r}$! At 8+ the results approximate to the tautology which in this situation surely reflects a complete lack of a system of counter drawing by the children. At 11+ disjunction is seen not in its inclusive form (*Either . . . or . . . or both*) but only as mutual exclusion (*Either . . . or*).

Lastly, we will demonstrate the value of the method in showing how far an instruction concerning reds is not taken specifically by younger children but is applied also to the other colours. A selection of the results of comparing the matching and non-matching of blue, yellow and green beads and counters is given in Table 10 for the games of implication and incompatibility.

Here is exhibited the matching spread of the implication rule in the case of younger children, from the reds to other colours, and to a slighter extent also the spread of the non-matching of incompatibility. By the age of 11+ the

TABLE 10. PROPORTIONS OF MATCHING AND NON-MATCHING OF NON-RED CONJUNCTIONS $\bar{R}\bar{r}$ BY CHRONOLOGICAL AGES

		Implication		Incompatibility	
		m	$\bar{m}$	m	$\bar{m}$
Expected proportions by chance		0·33	0·67	0·33	0·67
Age group	5+	0·91	0·09	0·21	0·79
	6+	0·59	0·41	0·24	0·76
	8+	0·38	0·62	0·18	0·82
	11+	0·38	0·62	0·34	0·66

spread of the rule seems to have disappeared, for the proportions are those expected by chance.

The above tables of results are given merely to illustrate the working of the theoretical analysis and the sensitivity and effectiveness of the method of experimentation. The full discussion of all the results against both chronological and mental ages will be given in a forthcoming monograph.

6. DISCUSSION

The purpose of this paper is to demonstrate a method and theoretical framework for investigating children's understanding of logical connectives used in binary propositions. We are not here concerned primarily with the results of the research or with their educational significance but, before passing on to a comment on the method and its underlying theory, some observations will be made on aspects of the results which have some methodological significance.

Taking first the *forward* games, the results of the implication game link closely with those obtained by Matalon (1962) who found that implication is predominantly read as equivalence. The difference found between the children's comprehension of implication and incompatibility (which is read by them as the logician defines it) appears to be new evidence. Disjunction caused most difficulty when the rule was expressed in the form: 'There must be at *least one red*.' Younger children were at a complete loss and the oldest children took the instruction to mean: 'There must be *only one red*.' They read disjunction in its exclusive sense as mutual exclusion.

The most noteworthy feature of the *reverse* games was the trend, more pronounced with older children, introduced when the experimenter made the valid responding draws of the counters which were not often made by the children in the forward games. Clearly the reverse game technique might have some value as a device for teaching propositions.

How far did the bead-counter game technique enable all the logical aspects of binary propositional thinking to be probed? Both Matalon (1962) and Hill (1961) were able to test the *modus tollendo tollens* logical situation. The form of their binary proposition tests involved the following arguments, Matalon using all four and Hill the extremes only, the *modus ponendo ponens* and *tollendo tollens:*

$$\begin{array}{c} p \supset q \\ p \\ \hline \therefore \text{What? } (q) \end{array}$$

modus ponendo ponens

$$\begin{array}{c} p \supset q \\ \bar{p} \\ \hline \therefore \text{What? } (q \vee \bar{q}) \end{array}$$

indeterminate

$$\begin{array}{c} p \supset q \\ q \\ \hline \therefore \text{What? } (p \vee \bar{p}) \end{array}$$

indeterminate

$$\begin{array}{c} p \supset q \\ \bar{q} \\ \hline \therefore \text{What? } (\bar{p}) \end{array}$$

modus tollendo tollens

It was not possible to test the *modus tollendo tollens* inference in the bead-counter game since it would have involved a non-red counter being drawn and a subsequent question being put to the child as to what kind of bead draw would have had to precede the counter draw. This would be calling for a certain abstracted quality of inference which it was not desired to invoke.

However, it would be quite possible to test the *modus tollens* inference of implication by having the child observe a bead-counter game between two other people and allowing him to see only what coloured counter the second player drew, and then finally asking him to say what the first player had drawn.

The disjunction inference, *modus tollendo ponens*, schematized as follows:

$$\begin{array}{c} p \vee q \\ \bar{p} \\ \hline \therefore \text{What? } (q) \end{array}$$

used by Hill (1961) was tested and found, as by her, to give no difficulty. A much more crucial test is the indeterminate disjunctive argument:

$$\begin{array}{c} p \vee q \\ p \\ \hline \therefore \text{What? } (\bar{q} \text{ or } q) \end{array}$$

As already mentioned, this showed an almost complete lack of q responses, indicating that *at least one* is taken to mean *one only*.

In general, the bead-counter game, as used to date, enables one to test inferences from the antecedent. If it were desired to investigate inferences from the consequence, as in the *modus tollendo tollens*, then the child could observe a game between two other people and then be asked for his inference concerning the initial bead draw.

We now turn to a fiducial question. The results of the games are compared with certain theoretical models based on propositional logic. Essentially obtained frequencies or proportions are being compared with those expected according to the logical definitions of the binary propositions. Thus in Table 5 the following frequencies for the 5+ group of children are being compared:

Conjunction	Rr	$R\bar{r}$
Theoretically expected frequencies	89	nil
Obtained frequencies	83	6

What confidence, statistically, have we that 83 can be regarded as an approximation to 89? It might appear to be a straight goodness of fit problem, comparing 83 and 6 against 89 and nil, but the nil entry rules out this approach. There is a further complication that each child makes 6 responses to R draws (and 4 to $\bar{R}$ draws). Should these 6 conjunctions be regarded as correlated? (An alternative method might be to allow each child only two responses, one to a first R draw and another to a first $\bar{R}$ draw. But even if this is done, the first difficulty of making a statistical comparison appears to remain.)

In this paper we have relied simply on the examination of proportions, as in Table 9 on the disjunction results, to arrive at a comparison and at some identification of the children's responses.

To sum up, the method enables one to pin-point precisely the particular conjunction (Rr, $R\bar{r}$, $\bar{R}r$ or $\bar{R}\bar{r}$) which are producing the general effects observed when a child is asked to interpret certain binary connectives. By the use of the proportions of the occurrences of the various conjunctions it is possible to make clear comparisons with whatever theoretical model is considered relevant. The games method also has real advantages as a psychological technique in holding the interest of the children and in rendering binary propositions acceptable to them.

Lastly, the method, particularly the reverse game technique, could be developed as a learning experiment to provide information about a more complex and significant type of human behaviour situation than those sometimes studied.

ACKNOWLEDGEMENTS

The actual experiments were carried out largely by my research assistant, Mrs. F. M. Davies, to whom credit for many of the details of experimental procedure and for recording of responses is due.

REFERENCES

HILL, S. A. (1961). *A study of the logical abilities of children.* Unpub. Ph.D. dissertation, Stanford University.

MATALON, B. (1962). Étude génétique de l'implication. *Études d'épistemologie génétique, XVI, Implication, formalisation et logique naturelle*, 69–93.

MAYS, W. (1963). Research on teaching primary school children to reason logically by using Lewis Carroll's *Game of Logic*. (An article appeared on this research in the *Times Educational Supplement*, p. 391, 27th September 1963.)

NAESS, A. (1962). L'emploi de la disjonction chez les adolescents. *Études d'épistemologie génétique, XVI, Implication, formalisation et logique naturelle*, 151–164.

SUPPES, P. (1965). On the behavioral foundations of mathematical concepts. *Child Dev. Mon.*, Ser. 99, **30**, No. 1, 89–93.

WILKINS, M. C. (1929). The effect of changed material on ability to do formal syllogistic reasoning. *Arch. Psychol.* No. 102.

Vol. 20
Part 1
93–106

The British Journal of Mathematical and
Statistical Psychology

May
1967

A METHOD OF CLUSTER ANALYSIS

By Paul Constantinescu
University of Bucharest

A mathematical method is proposed for clustering a set of elements with respect to a set of criteria. Input information is described in terms of a graph representing a distance function between the elements. Definitions are given for the concepts of a ' tree ', and a ' tree-representation '. It is shown that clusters are particular kinds of subsets (level-subsets) in a tree-representation; and that a final pattern contains at least one non-trivial tree, under specified conditions. An example of clustering is set out, and the computation of level-subsets is described. The selection of clusters from a final pattern is briefly discussed.

1. Introduction

Although the problems associated with the classification of a set of elements according to some criteria have been considered for a long time, it is only recently in the era of high-speed computers (Ball, 1965) that practical solutions have become available. The increasing quantities of data that have to be analysed in many fields of science demand methods for clustering the elements involved in order to simplify the analysis to a point where only the ' representatives ' of the clusters have to be studied. This process is already a classic one in mathematics where clustering takes place for sets endowed with different algebraical or topological structures. Empirical problems of taxonomy generally have to be examined with sets which do not have such properties; the criteria, in respect of which the classification has to be made, yield a distance or similarity function defined for the set of all pairs of elements. The starting point for most clustering problems may be considered to be a matrix $\mathbf{M}_d$ which describes the distances (or similarities), whose order is the number of elements to be clustered and whose entries are real numbers associated with the pairs of elements.

2. The Cluster

From the above viewpoint, two ways of approaching a mathematical definition of a cluster suggest themselves. One of these conceptions of a cluster which is relevant to applications in psychology and biology will now be described.

For the sake of simplicity it will be assumed that the numbers associated with the pairs of elements (the entries in $\mathbf{M}_d$) are non-negative integers 0, 1, 2, . . ., N. If the semi-circles $C_1, C_2, \ldots, C_N$ of radius $1, 2, \ldots, N$ are represented on the half-plane, then a subset $E_1, E_2, \ldots, E_q$ of the set E of

m elements will be in C_k (and not in C_{k-1}), if and only if all the $q(q-1)/2$ pairs which can be formed with these q elements are associated with entries in $\mathbf{M}_d$ which are less than or equal to k, and if for at least one pair the entry is k (Figure 1).

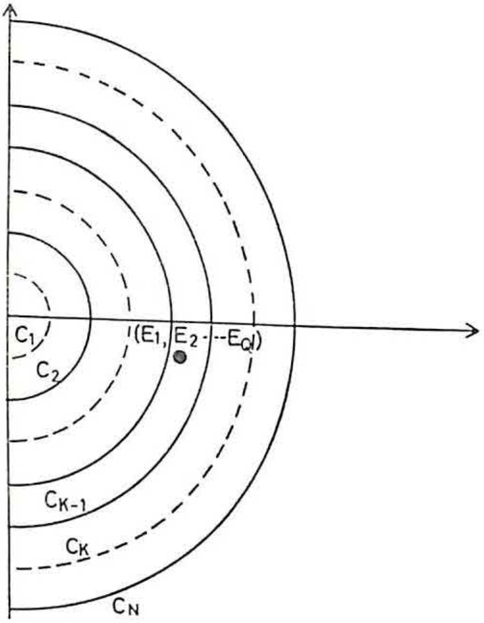


FIGURE 1. The representation of subsets within semi-circles on the half-plane. $E_1, E_2, \ldots, E_q$: a subset of the set E of m elements. $C_1, C_2, \ldots, C_N$: semi-circles of radius 1, 2, . . . N on the half-plane.

A system of subsets of E is chosen which fulfils the following conditions:

(*a*) the set E which is in C_N belongs to the system;

(*b*) only disjoint subsets which are in C_k ($k=0, 1, \ldots, N$) may be considered in the system;

(*c*) if the subsets S and S^* are in C_p and C_q respectively, where $p<q$, then either $S \subset S^*$ or $S \cup S^* = \phi$.

The subsets of the system subject to conditions (*a*), (*b*) and (*c*) will be called 'clusters'.

A second approach, which seems to be useful for some technological problems (for instance, the clustering of the vertices of a graph as a step in the logical design of a computer) is that a cluster must be produced only by all the pairs of the elements whose associated entries in $\mathbf{M}_d$ are exactly equal to k.

This paper deals with the first case, and formalizes the concept of 'cluster' in this sense. It is stressed that the system of clusters which fulfils conditions (*a*), (*b*) and (*c*) is not a partition of the set E as required, for instance, by Regnier

(1965); the union of the subsets is the set E, but according to (*c*) the subsets of the system are not disjoint.

A formal mathematical scheme to solve the clustering problem in this sense, already briefly described by Constantinescu (1966), will be presented; and its relation to the problem solved by Regnier (1965) will be briefly examined. Computing programs in Algol[1], embodying different variants of the algorithm, have been written and an application of the clustering method in psychology is described by Stringer (1967).

3. THE INPUT INFORMATION

Let $E=\{E_1, E_2, \ldots, E_m\}$ be a set of elements, and $P=\{P_1, P_2, \ldots, P_n\}$ a set of criteria. The criteria engender the function.

$$f : E \rightarrow p_1 \times p_2 \times \ldots \times p_n$$

where p_l $(l=1, 2, \ldots, n)$ is the set of values associated with the criteria P_l.

The function f can be described by a matrix $\mathbf{M}_0$ whose row vectors $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_m$ are in one-to-one correspondence with $E_1, E_2, \ldots, E_m$.

For instance, if all the p_l are sets $\{0, 1\}$ the classification problem is a binary one for which the row vectors are in a Euclidean space of dimension n over the field $\{0, 1\}$.

Since the elements of the p_l are real numbers a distance function can be associated with the elements of E,

$$d : E \times E \rightarrow I,$$

where $I=\{0, 1, 2, \ldots, N\}$. The values of d can always be considered as non-negative integers, although it will sometimes be necessary to add a constant to the entries in $\mathbf{M}_0$ or multiply them by a constant; or make similar adjustments to the symmetrical distance matrix $\mathbf{M}_d=(d_{ij})$, where $d_{ij}=d(\mathbf{e}_i, \mathbf{e}_j)$ and $d_{ii}=0$ $(i=1, 2, \ldots, m)$.

The function d can also be described by a graph G (Figure 2), as follows. With every element of I a level is associated. On level 0 the elements of E are considered as vertices. On level k, vertices are formed by drawing lines from level 0 for pairs (E_i, E_j) for which $d_{ij}=k$. The vertices are marked with the appropriate (E_i, E_j). This graph will be called the 'graph of pairs'.

If the lines are not drawn, the pattern G' so obtained from G will be called the 'representation G'' of the graph G, or the 'representation of the pairs'.

In G (or G') a vertex can be considered, marked by the subset $E_{i_1}, E_{i_2}, \ldots E_{i_q}$ on level k $(0 \leqslant k \leqslant N)$, if $d_{i_s i_t} \leqslant k$ for all $s, t=1, 2, \ldots, q$, and if there are $s', t' \in \{1, 2, \ldots, q\}$ such that $d_{i_{s'} i_{t'}}=k$ (Figure 3). The method for constructing such subsets (or vertices) will be given below. These subsets will be called '*level-subsets*'.

---

[1] Programs are available at the SRC Atlas Computer Laboratory, and at the Computing Centre of the University of Bucharest.

The vertices $E_1, E_2, \ldots, E_m$ on level 0 and the pairs of elements are particular level-subsets.

The pattern (Figure 3) in which the level-subsets $\{E_{i_1}, E_{i_2}, \ldots, E_{i_q}\}$ are displayed with respect to the distance function d will be called the '*final pattern*' with respect to d. Different systems of clusters can be selected from it, fulfilling conditions (a), (b) and (c) of Section 1.

In brief, the input information for the clustering problem comprises the matrix $\mathbf{M}_d$, or the graph G, or the representation G' of the graph G. Each of these contains the same amount of information.

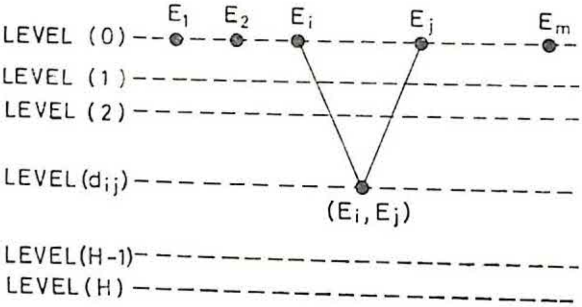


FIGURE 2. The graph and the representation of pairs
$E_1, E_2, \ldots, E_m$: elements of the set E considered as vertices.
E_i, E_j: pairs linked at the appropriate distance level.

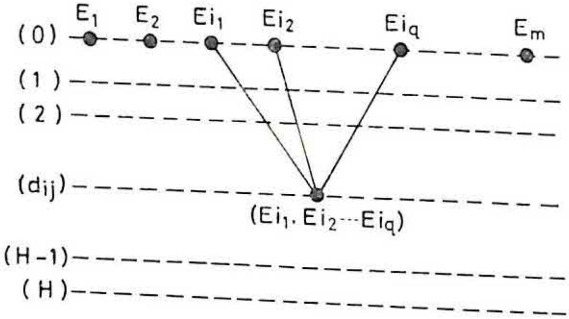


FIGURE 3. A level-subset marked by a vertex in a graph.
$E_1, E_2, \ldots, E_m$: elements of the set E.
$E_{i_1}, E_{i_2}, \ldots, E_{i_q}$: level-subsets or vertices.

4. DUALITY IN CLUSTERING PROBLEMS

Having defined the function d, a new function $s : E \times E \to I$ can be defined by

$$s(\mathbf{e}_i, \mathbf{e}_j) = s_{ij} = N - d_{ij}.$$

The function s is the similarity of the vectors $\mathbf{e}_i$, $\mathbf{e}_j$. The definitions with respect to s of the matrix M_s, the graph G, the representation G' and the final pattern follow as for d. Similarity and distance are dual in the sense of the following lemma:

Lemma 1

If $\{E_1, E_2, \ldots, E_q\}$ is a subset on level k in the final pattern with respect to d, then $\{E_1, E_2, \ldots, E_q\}$ is a subset on level $N-k$ in the final pattern with respect to s, and reciprocally.

Proof

If the subset $\{E_1, E_2, \ldots, E_q\}$ is on level k, then all the $q(q-1)/2$ pairs are on the first k levels in G: that is:

$$d_{ij} \leqslant k \text{ for all } i, j \text{ for which } E_i, E_j \in \{E_1, E_2, \ldots, E_q\};$$
$$-d_{ij} \geqslant -k \text{ for all } i, j \text{ for which } E_i, E_j \in \{E_1, E_2, \ldots, E_q\};$$
$$N-d_{ij} \geqslant N-k \text{ for all } i, j \text{ for which } E_i, E_j \in \{E_1, E_2 \ldots, E_q\};$$
$$s_{ij} \geqslant N-k \text{ for all } i, j \text{ for which } E_i, E_j \in \{E_1, E_2, \ldots, E_q\};$$

and there are $i', j' \in \{1, 2, \ldots, q\}$ such that $d_{i'j'} = k$. Hence there are $i', j' \in \{1, 2, \ldots, q\}$ such that $s_{i'j'} = N-k$, then $E_1, E_2, \ldots, E_q$ is a subset on level $N-k$ on the final pattern with respect to s. For the reciprocal situation the proof is similar.

On the basis of this lemma, clustering algorithms can be elaborated with respect to either d or s, and the results used by duality in the other case.

5. THE TREE AND THE TREE REPRESENTATION

Definitions

(I) The level p is *inferior* to the level q if $p \geqslant q$.

(II) The vertices on the level 0 will be called '*initial vertices*' and the vertices on the last level (N) will be called '*final vertices*'.

(III) '*Absorption*' is the operation by which, from any level-subsets and pairs of elements, another level-subset is obtained which is equal to their union and placed on the most inferior of the levels on which these pairs or level-subsets are placed. This operation is permissible if and only if, in the pairs and level-subsets which make up the new level-subset, all the pairs which can be formed with the elements of their union are present.

Any of the pairs and level-subsets which make up the level-subset union may or may not be included in the new pattern in which the level-subset union is included. Those which are not represented are said to be absorbed by the level-subset union.

The final pattern in which the level-subsets are represented, and from which different systems of clusters can be selected according to (*a*), (*b*) and (*c*) of Section 1, is obtained from the graph of pairs (representation of pairs) by absorptions.

By absorptions a final vertex representing the whole set E is obtained on the last level.

(IV) A graph G, in which vertices marked by level subsets are displayed, will be called a '*tree*' if

(i) it has only one final vertex, and

(ii) one and only one branch goes out to the inferior levels from every vertex, including the initial vertices.

According to the definitions of the graph of pairs and of absorption, a branch will be drawn between two vertices S and S^*, if and only if $S \subset S^*$ and there is no vertex on an intermediate level representing a subset T such that $S \subset T$.

Different trees can be obtained from the graph of pairs by absorptions.

(V) A representation G' in which level-subsets are represented will be called a 'tree-representation', if

(i) it has only one final level-subset,

(ii) on each level the level-subsets are disjoint, and

(iii) a level-subset S on level $s (s \neq 0)$ contains at least one element of the subset R on level $r < s$, such that $R \subset S$.

Different tree-representations can, by absorption, be obtained from the representation of pairs.

Lemma 2

The level-subsets in a tree-representation (including pairs of elements or any elements on level 0) are clusters, and conversely.

Proof

(1). The level-subsets in a tree-representation are clusters. Proposition (*a*) of Section 1 follows from V (i). Since there is only one final level-subset (on the last level) it must be the whole set E. Hence the set E belongs to the system.

From V (ii) follows (*b*) of Section 1. From V (iii) it follows that for the level-subsets S and S^* in C_p and C_q respectively $(p < q)$, there are only two possibilities:

$$S \subset S^*$$
$$S \cap S^* = \phi.$$

(2). The clusters are level-subsets in a tree-representation.

From (*a*) follows V (i). If there were another final level-subset, it would be absorbed by the final level-subset E.

From (*b*) follows V (ii). On each level the level-subsets are disjoint.

From (*c*) follows V (iii). If the level-subsets S and R on level s and $r (r < s)$ respectively are not disjoint—that is, S contains at least one element of the subset R—then $R \subset S$.

THEOREM 1

G' is a tree-representation if and only if G is a tree. By erasing the branches from a tree G, a tree-representation G' is obtained; by drawing the branches in a tree-representation G' a tree G is obtained.

Proof

(1). Let G be a tree. By erasing the branches a system of level-subsets with the properties V (i), V (ii) and V (iii) is obtained. Clearly V (i) follows, namely that there is only one final level-subset.

It will now be shown by negation that the level-subsets on each level are disjoint.

Let there be two level-subsets S and S^* on level r, such that $N=S\cap S^*$ and $N\neq\phi$. Let $E_i\in N$. There is a subset M on level $p<r$ such that $E_i\in M\subset N\subset S$ (finally $M=\{E_i\}$ will be on level 0), and also a subset M^* on level $q<r$ such that $E_i\in M^*\subset N\subset S^*$. If $M=M^*$ (in this case $p=q$), two branches go out from the same vertex to S and S^*. This contradicts V (ii). If $M\neq M^*$ the process is continued using M as S and M^* as S^*; finally stopping when $p=q=o$ and $M=M^*=\{E_i\}$. Hence V (ii) follows.

The argument by negation also shows that V (iii) follows. If $E_i\in R$ and $E_i\in S$ and $R\not\subseteq S$, then a branch cannot be drawn from vertex R to vertex S. This contradicts IV (ii).

(2). Let G' be a tree-representation. V (i) implies IV (i). To prove that IV (ii) follows from V (ii) and V (iii) it will be recalled that a branch from S on level r is drawn to S^* on level s (where $r<s$), if $S\subset S^*$ and there is no T on level p (where $r<p<s$) such that $S\subset T\subset S^*$. Let S be a level-subset on level r. All the level-subsets on inferior levels are, according to V (iii), disjoint with S, or some of them include S. At least one includes S (for instance, E on the last level). Let all the subsets which include S be

$$\begin{array}{l} S_1 \text{ on level } i_1 \\ S_2 \text{ on level } i_2 \\ \cdot \\ \cdot \\ \cdot \\ S_t \text{ on level } i_t, \end{array}$$

where $r<i_1<i_2\ldots i_t$. $S\subset S_k$ for all k for which $1\leqslant k\leqslant t$.

By V (iii):

$$S\subset S_1;\quad S_1\subset S_2;\quad \ldots;\quad S_{t-1}\subset S_t.$$

Since $S\subset S_1\subset S_2\ldots\subset S_t$, one and only one branch goes out from S to S_1. (Between levels r and i_1 there is no intermediate level on which there is an S^* such that $S\subset S^*$.) Hence IV (ii) follows.

Corollary

The vertices in a tree are clusters, and conversely.

6. THE EXISTENCE THEOREM

A tree (tree-representation) will be called *trivial* if it has exactly $m+1$ vertices (m is the number of elements in the set E), namely one final vertex and m initial vertices (Figure 4).

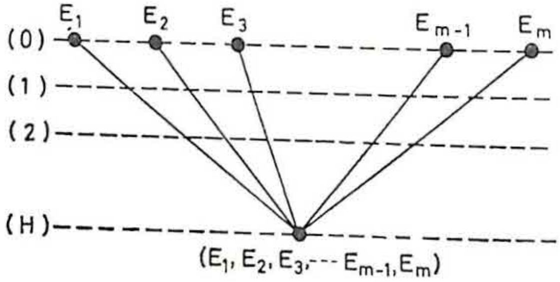


FIGURE 4. The trivial tree-representation.
$E_1, E_2, \ldots, E_m$: elements of the set m.

THEOREM 2

A final pattern contains at least one non-trivial tree-representation if and only if there are k level-subsets with the properties:

(i) $1<k<m$,

(ii) at least one level-subset is not on the last level, and

(iii) the k level-subsets represent a partition for E.

Proof

(1). Let there be a final pattern in which there are k level-subsets with the properties (i), (ii) and (iii). It can be deduced from (i) that these subsets do not form the partition $E=\bigcup_i\{E_i\}$ or the partition $E=E$.

From this pattern the following tree-representations may be obtained:

$\mathscr{S}_1$, the set E on the final level;

$\mathscr{S}_2$, the initial subsets on level 0;
$\{E_1\}, \{E_2\}, \ldots, \{E_m\}$, and

$\mathscr{S}_3$, the k level-subsets.

From property (i) follows V (i); all the other subsets on the last level are absorbed by E. Since the k subsets are disjoint, then the subsets are also disjoint on each individual level; and the initial subsets are also disjoint. Hence V (ii) follows. The proposition V (iii) follows from the fact that inclusions are only permissible between the initial subsets and the k subsets and E.

(2). Let there be a non-trivial tree-representation in the final pattern. There are k subsets with the properties (i), (ii) and (iii) of Theorem 2. Because the tree-representation is non-trivial there is at least one level-subset S on a level s (where $1 < s < N$). The k level-subsets are the subsets S which are disjoint together with the subsets $\{E_i\}$ which are needed to form by union the set E.

It follows that $1 < k < m$ and the properties (ii) and (iii) are fulfilled.

Corollary

From Theorems 1 and 2 it follows that a final pattern contains at least one non-trivial tree, if and only if the conditions of Theorem 2 are fulfilled.

7. COMPUTING LEVEL-SUBSETS

Before the computation is described an example will be given.

Assume the following input data:

$$\mathbf{M}_0 = \begin{array}{r|ccccccccccccc}
1 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
2 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 1 \\
3 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\
4 & 1 & 1 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\
5 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\
6 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 1 \\
7 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
8 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 1 & 1 & 0 \\
9 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\
10 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0
\end{array}$$

The matrix $\mathbf{M}_d$ of distances between elements is:

$$\begin{array}{cccccccccc}
\mathbf{1} & \mathbf{2} & \mathbf{3} & \mathbf{4} & \mathbf{5} & \mathbf{6} & \mathbf{7} & \mathbf{8} & \mathbf{9} & \mathbf{10} \\
0 & 3 & 4 & 4 & 2 & 7 & 7 & 9 & 1 & 6 \\
 & 0 & 5 & 3 & 3 & 10 & 10 & 6 & 2 & 9 \\
 & & 0 & 4 & 6 & 11 & 7 & 5 & 5 & 8 \\
 & & & 0 & 6 & 9 & 9 & 7 & 5 & 6 \\
 & & & & 0 & 7 & 7 & 9 & 1 & 6 \\
 & & & & & 0 & 4 & 10 & 8 & 3 \\
 & & & & & & 0 & 6 & 8 & 5 \\
 & & & & & & & 0 & 8 & 11 \\
 & & & & & & & & 0 & 7 \\
 & & & & & & & & & 0
\end{array}$$

The representation of pairs is:

1			$\overset{\cdot}{(1,9)}$		$\overset{\cdot}{(5,9)}$		
2			$\overset{\cdot}{(1,5)}$		$\overset{\cdot}{(2,9)}$		
3		$\overset{\cdot}{(1,2)}$	$\overset{\cdot}{(2,4)}$	$\overset{\cdot}{(2,5)}$	$\overset{\cdot}{(6,10)}$		
4		$\overset{\cdot}{(1,3)}$	$\overset{\cdot}{(1,4)}$	$\overset{\cdot}{(3,4)}$	$\overset{\cdot}{(6,7)}$		
5		$\overset{\cdot}{(2,3)}$	$\overset{\cdot}{(3,8)}$	$\overset{\cdot}{(3,9)}$	$\overset{\cdot}{(4,9)}$	$\overset{\cdot}{(7,10)}$	
6	$\overset{\cdot}{(1,10)}$	$\overset{\cdot}{(2,8)}$	$\overset{\cdot}{(3,5)}$	$\overset{\cdot}{(4,5)}$	$\overset{\cdot}{(4,10)}$	$\overset{\cdot}{(5,10)}$	$\overset{\cdot}{(7,8)}$
7	$\overset{\cdot}{(1,6)}$	$\overset{\cdot}{(1,7)}$	$\overset{\cdot}{(3,7)}$	$\overset{\cdot}{(4,8)}$	$\overset{\cdot}{(5,6)}$	$\overset{\cdot}{(5,7)}$	$\overset{\cdot}{(9,10)}$
8		$\overset{\cdot}{(3,10)}$	$\overset{\cdot}{(6,9)}$	$\overset{\cdot}{(7,9)}$	$\overset{\cdot}{(8,9)}$		
9		$\overset{\cdot}{(1,8)}$	$\overset{\cdot}{(2,10)}$	$\overset{\cdot}{(4,6)}$	$\overset{\cdot}{(4,7)}$	$\overset{\cdot}{(5,8)}$	
10			$\overset{\cdot}{(2,6)}$	$\overset{\cdot}{(2,7)}$	$\overset{\cdot}{(6,8)}$		
11				$\overset{\cdot}{(3,6)}$	$\overset{\cdot}{(8,10)}$		

The representation of pairs is replaced by two arrays $\mathbf{G}_1$ and $\mathbf{G}_2$, whose rows are in one-to-one correspondence with the levels. The indices which occupy the first position of the pairs on a given level are written in natural order on the corresponding row of $\mathbf{G}_1$; the corresponding indices for the second position are written on each row of $\mathbf{G}_2$. In the present example:

$\mathbf{G}_1$:
1 5
1 2
1 2 2 6
1 1 3 6
2 3 3 4 7
1 2 3 4 4 5 7
1 1 3 4 5 5 9
3 6 7 8
1 2 4 4 5
2 2 6
3 8

$\mathbf{G}_2$:
9 9
5 9
2 4 5 10
3 4 4 7
3 8 9 9 10
10 8 5 5 10 10 8
6 7 7 8 6 7 10
10 9 9 9
8 10 6 7 8
6 7 8
6 10

The computation of the level-subsets is effected by successively exploring the rows of these arrays by means of a variable $b = 1, 2, \ldots, 11$ and by applying the computer procedure TRIANG. By this procedure the level-subsets are constructed on each level, starting with the representation (or graph) of pairs.

For each value of b two arrays $\mathbf{T}_1$ (an $N \times 1$ column vector) and $\mathbf{T}_2$ (an $N \times f$ matrix) are constructed (where f is the maximum number of indices which are written on a line in $\mathbf{G}_1$ or $\mathbf{G}_2$. All the distinct numbers which can be found on the first b rows in $\mathbf{G}_1$ are written in natural order in $\mathbf{T}_1$. All the elements from the first b rows of $\mathbf{G}_2$ which are in correspondence with the respective numbers from $\mathbf{T}_1$ are written in natural order in the corresponding rows of $\mathbf{T}_2$. All the other elements in $\mathbf{T}_1$ and $\mathbf{T}_2$ are 0.

In the example, if $b = 3$, then:

$$\mathbf{T}_1 = \begin{array}{c} 1 \\ 2 \\ 5 \\ 6 \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \end{array} \qquad \mathbf{T}_2 = \begin{array}{ccccccc} 2 & 5 & 9 & 0 & 0 & 0 & 0 \\ 4 & 5 & 9 & 0 & 0 & 0 & 0 \\ 9 & 0 & 0 & 0 & 0 & 0 & 0 \\ 10 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & & & & & & 0 \\ \cdot & & & & & & \\ \cdot & & & & & & \\ \cdot & & & & & & \\ 0 & & & & & & 0 \end{array}$$

In order to determine the level-subsets for level b, the square matrix, $\mathbf{R}$, is formed from $\mathbf{T}_2$ which for each pair of row and column has at least one element different from zero. In the present example, with $b = 3$, the resulting square matrix is:

$$\mathbf{R} = \begin{array}{cccc} 2 & 5 & 9 & 0 \\ 4 & 5 & 9 & 0 \\ 9 & 0 & 0 & 0 \\ 10 & 0 & 0 & 0 \end{array}$$

Then from $\mathbf{T}_2$, all the triangular arrays $\mathbf{U}_2$ are formed (with the elements in natural order) such that the elements which lie on the second diagonal are all equal to one of the members of the second diagonal of the square matrix $\mathbf{R}$. For each array there is associated a column vector $\mathbf{U}_1$ composed of the elements from $\mathbf{T}_1$ which correspond to the elements of $\mathbf{T}_2$ included in $\mathbf{U}_2$. So for level b all arrays $\mathbf{U}_1$ and $\mathbf{U}_2$ of the following kind are found:

$$\mathbf{U}_1 = \begin{array}{c} i_1 \\ i_2 \\ i_3 \\ \cdot \\ i_{k-1} \end{array} \qquad \mathbf{U}_2 = \begin{array}{lllll} i_2 & i_3 & i_4 & \ldots & i_k \\ i_3 & i_4 & \ldots & i_k & \\ i_4 & \ldots & i_k & & \\ \cdot & & & & \\ i_k & & & & \end{array}$$

In the present example the arrays for $b=3$ are:

$\mathbf{U}_1$	$\mathbf{U}_2$
1	2 5 9
2	5 9
5	9
6	10
1	5 9
5	9

The following theorem will now be proved.

THEOREM 3

If a set $\{i_1, i_2, \ldots, i_k\}$ obtained from a given $\mathbf{U}_1$ and $\mathbf{U}_2$ is not included in a level-subset on a superior level (including b) then it is a level-subset on level b.

Proof

$d_{i_s i_t} \leqslant b$, since $i_s, i_t \in \{i_1, i_2, \ldots, i_k\}$. Because the set is not included in a level-subset on a superior level there is at least one pair s', t' such that $d_{i_{s'} i_{t'}} = b$. All the $k(k-1)/2$ pairs which can be formed with $i_1, i_2, \ldots, i_k$ are present in $\mathbf{U}_1$ and $\mathbf{U}_2$. The latter contain distinct pairs; and only pairs formed with elements from $i_1, i_2, \ldots, i_k$, i.e. exactly $1+2+3+\ldots(k-1) = k(k-1)/2$.

In the example, the sets (1, 2, 5, 9), (6, 10), (1, 5, 9) are obtained. The set (1, 5, 9) is not a level 3 subset, because it will be found to appear on level 2. (In order that only those subsets which are not included on a superior level are written on any level, a computer procedure INCLUD is used).

In the example this leads to the following subsets:

1 (1,9) (5,9)
2 (1,5,9) (2,9)
3 (1,2,5,9) (6,10)
4 (1,3) (3,6) (6,7)
5 (1,2,3,4,9) (6,7,10)
6 (1,2,3,4,5,9) (1,10) (4,5,10) (7,8)
7 (1,4,5,9,10) (1,5,6,7,10) (3,5,7)
8 (1,3,4,5,9,10) (1,5,6,7,9,10) (2,8,9) (2,3,8,9)
(3,5,7,9,10) (3,7,8,9)
9 (1,2,3,4,5,8,9) (1,3,4,5,7,8,9) (1,4,5,6,7,9,10) (1,2,9,10)
(1,2,3,9,10) (1,2,3,4,9,10) (1,2,3,4,5,9,10) (3,4,5,7,9,10)
10 (1,2,3,4,5,7,8,9) (1,4,5,6,7,8,9) (1,2,4,5,6,7,8,9)
(1,2,3,4,5,7,9,10)
11 (1,2,3,4,5,6,7,8,9,10)

By slightly modifying the algorithm, it is possible to exclude those level-subsets which are contained in level-subsets which are situated together on any lower level. The result is the final pattern from which different systems of clusters can be selected by absorptions according to (*a*), (*b*) and (*c*) of Section 1. From the final pattern of the example, the trees in Figure 5 can, for instance, be selected.

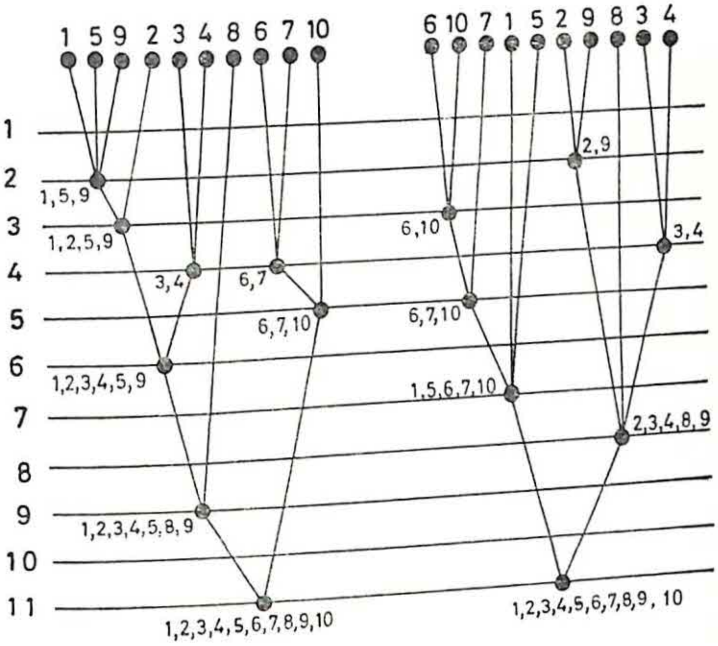


FIGURE 5. Trees in a final pattern.

8. COMPUTING THE CLUSTERS

If only those subsets are written on every level which are disjoint with the level-subsets already written on the same level, and are disjoint with or include the level-subsets on preceding levels, then a tree is directly obtained in the final pattern. By permutation of the level-subsets it is possible to find all the trees that may be associated with any given classification problem.

But for many problems in psychology and biology it seems to be more useful to produce a final pattern from which many different trees can be selected. The final problem then concerns the number of level-subsets producing the clusters, and implicitly the number of trees; that is, the number of classification possibilities. To determine this complexity of the final pattern of level-subsets five alternative algorithm have been briefly described elsewhere (Constantinescu, 1966); other variants are also available. If the clustering method described by Regnier (1965) is called the 'one-level clustering method', the method presented here may be called a 'multi-levels clustering method'.

ACKNOWLEDGEMENTS

The computing programs referred to in this paper were written during 1964–65 when the author was at the SRC Atlas Computer Laboratory. I am indebted to Dr. J. Howlett and Dr. R. Churchouse for their assistance and encouragement during this visit.

I should like to thank Professor R. J. Audley and Mr. P. Stringer for criticism and help in writing this paper.

REFERENCES

BALL, G. H. (1965). Data analysis in the social sciences: What about the details? *Fall Joint Computer Conference. AFIPS Conference Proceedings*, **27**, 533–559. Washington, D.C.: Spartan Books.

CONSTANTINESCU, P. (1966). The classification of a set of elements with respect to a set of properties. *Computer Journal*, **8**, 352–357.

REGNIER, S. (1965). Sur quelques aspects mathématiques des problèmes de classification automatique. *ICC Bulletin*, **4**, 175–191.

STRINGER, P. (1967). Cluster analysis of non-verbal judgements of facial expressions. *Br. J. math. statist. Psychol.*, **20**, 71–79.

NOTES AND CORRESPONDENCE

COMPARISON OF FACTORS BY AHMAVAARA'S METHOD

By Max Hamilton
University of Leeds

The comparability of factors obtained from different sets of data has always been a problem of interest but great difficulty to psychologists. A number of methods have been proposed, but of them Ahmavaara's (1954) is the simplest, neatest and most elegant. This may fairly safely be used when the variables in the populations have equal variance. Despite this it has been little used and rarely referred to. In the book by Harman (1960) there is a brief reference to the work of Ahmavaara. The method of comparing different factors is described by Fruchter and Jennings (1962), but unfortunately there is an error in the description. For this reason an account of the method is thought to be worth publishing in relation to a practical example. The basic theorems and the proofs of the formulae are not easy, but the following simplified account does give a useful introduction to the method.

The data in this example is based on ratings of 152 men and 120 women suffering from primary depressive illness, using a 17-item rating scale (Hamilton, 1960). From the ratings for the two sexes separately, two matrices of intercorrelation were calculated and then factor analysed by the method of principal components. These factors were then rotated by the Varimax method (Kaiser, 1958). Conforming to the criterion of Kaiser (1960) six factors were extracted. (The correlation matrices and factor saturations have been presented elsewhere (Hamilton, 1967)).

The six (orthogonal) factors constitute the basis of a six-dimensional subspace of the total space of 17 dimensions. Since the variables used for the rating of the men and the women are the same, the two sets of factors define two subsets in the same total space. It may therefore be possible to rotate one set of axes (factors) to occupy roughly the same position as the second set by means of a suitable transformation matrix (provided they occupy the same subspace). Such a transformation matrix will show the second set of factors as a linear combination (weighted sum) of the first set. It is obvious that if the two sets of factors were initially identical, the transformation matrix required to rotate one into the other would be the identity or unit matrix (provided that the two sets of factors are in the same order. If the order is not the same, the unit matrix will have rows or columns displaced). If the two sets of factors are not quite identical, then the transformation matrix will be not quite a unit matrix. Thus the way in which the transformation matrix resembles, or differs from, a unit matrix will correspond to resemblance or difference between the two sets of factors. Even more important, the transformation matrix will show the constitution of the second set in terms of the first set.

Given that the two sets of factors are represented by X and Y, with the rows representing the variables and the columns the factors (there being therefore less columns than rows), then the transformation matrix proposed by Ahmavaara is

$$L=(X'X)^{-1}X'Y$$

where X' is the transpose of X. Examination of the formula for L shows that when X and Y are identically equal, then L is a unit matrix. Since it is customary for variables to be represented by vectors of unit length, the rows of L should be normalized, i.e. the sums of the squares of the elements of each row should sum to unity. This is represented by L_n. The columns of L represent the common factors of the matrix Y and the rows that of X. Therefore each row of L_n represents the projection (loading) of the factor in X on the factors of Y. Ahmavaara's own words cannot be bettered (1954, p. 56) when describing L_n: "its columns represent the common factors of the matrix Y, the order of the factors being the same in L_n as in Y. On the other hand, the rows of L_n represent the common factors of X, the order of the factors being again the same in L_n as in X. The elements of L_n then indicate the cosines of the angles between the factors of the different studies or, more accurately, between the vectors representing these factors. Consequently, if some element of L_n is exactly equal to unity, the respective factors of the two studies are identical ". The correlations between the factors in the two sets can be evaluated by calculating L_nL_n'. Since L is the matrix for rotating X into Y, then the product XL should correspond closely to Y. This serves as a check on effectiveness of the comparison. Finally, it must be remembered that since L is not an orthogonal matrix, the rotated factors are not orthogonal. Should orthogonality be desired, a suitable method is described in Lawley and Maxwell (1964).

In the case of oblique factors the procedure is a little more complicated. To ensure that the vector is of unit length, the normalization must be carried out before the oblique rotation, for only with orthogonal axes does the sum of squares come to unity (extension of Pythagoras' theorem). Thus if F represents the matrix of factor saturations, the factors being orthogonal and obtained by any method, Λ represents the non-orthogonal rotation matrix and Y is the matrix of factor saturations in the rotated oblique factors, then

$$\text{for factors } X\text{:}\quad F_x\Lambda_x = V_x$$

$$\text{and for factors } Y\text{:}\quad F_{\mathrm{y}}\Lambda_{\mathrm{y}} = V_{\mathrm{y}}.$$

The initial transformation matrix $L=(V_x'V_x)^{-1}V_x'F_{\mathrm{y}}$. When the rows of L are normalized, then L_n is the transformation matrix which compares V_x with F_y. Then (quoting again) " when L_n is postmultiplied by Λ_y the results $L_n\Lambda_y$ is the matrix of comparison between the primary factors of X and the reference factors of Y. To be able to compare the primary factors with each other, each column has to be multiplied by a constant, obtained by calculating $(\Lambda_y'\Lambda_y)^{-1}$, then taking the diagonal elements as a diagonal matrix D^2, and postmultiplying the transformation matrix by D^{-1}, thus: $L_n\Lambda_yD^{-1}$."

The procedure can be applied for transforming Y into X. When the variables are not the same in the two sets of factors, the method can be applied to those variables that are common, ignoring the others.

Comparison of Factors in Depression

As a guide, the method was first applied to the principal components. The results (Table 1) show that the first factor for women (WF_1) corresponds almost exactly to the first factor for men (MF_1). There is very good agreement between WF_2 and MF_2. There is fair correspondence between WF_4 and MF_3 (but the scores run in the reverse direction) and for WF_3 and MF_4. WF_6 has a fair correspondence with MF_5 but an even greater one with MF_1. This leaves WF_5 and MF_6 as the remaining unmatched pairs, but in fact WF_5 corresponds more closely to MF_3 and MF_4 than to anything else. The general conclusion to be drawn from examination of these results is that the first four factors obtained from the matrix show reasonable resemblance between the two sexes. The lack of adequate correspondence between the last two might be put down to instability. An alternative suggestion is that the matrix contains only four factors worth considering.

TABLE 1. COMPARISON OF FACTORS BY AHMAVAARA'S METHOD: PRINCIPAL COMPONENTS

		From 152 men					
		1	2	3	4	5	6
From 120 women	1	1·00	−0·03	0·08	0·04	0·00	−0·02
	2	0·04	0·93	−0·00	0·28	−0·05	−0·24
	4	0·11	0·46	−0·70	−0·12	0·33	0·41
	3	0·01	−0·22	0·46	0·68	0·48	0·22
	6	0·65	0·08	−0·41	−0·34	−0·51	−0·16
	5	−0·06	−0·39	−0·66	0·59	−0·24	−0·03

TABLE 2. COMPARISON OF FACTORS BY AHMAVAARA'S METHOD: VARIMAX ROTATION

		From 152 men					
		1	2	3	4	5	6
	1	0·97	0·06	0·03	−0·12	0·11	−0·17
	2	−0·21	0·93	−0·15	0·00	0·13	0·20
	6	−0·01	0·24	−0·66	−0·30	−0·53	−0·36
From 120 women	5	0·01	−0·26	−0·25	0·92	−0·04	0·16
	4	0·10	−0·01	−0·29	−0·05	−0·74	0·59
	3	0·49	0·35	0·47	−0·14	0·08	0·62

Comparison of the rotated factors (Table 2) confirms the belief that the method of rotation tends to show the presence of invariant factor structure. The first factor of the women WV_1 corresponds very closely with that of the men MV_1. There is very good correspondence between WV_2 and MV_2, and also between WV_5 and MV_4. There is fair correspondence between WV_4 and MV_5, but the former contains also a large sized component from MV_6. WV_6 corresponds most closely to MV_3 but has a large component from MV_5 and similar ones from MV_4 and MV_6. Finally, WV_3 corresponds most closely to MV_6 but contains large components from MV_1 and MV_3. Since fair correspondence between all six factors has been obtained, the conclusion is that the possibility of instability of the two smallest principal components is a reasonable one.

TABLE 3. COMPARISON OF FACTORS BY AHMAVAARA'S METHOD: VARIMAX ROTATION

		From 70 women					
		1	2	3	4	5	6
	1	0·76	−0·49	−0·37	−0·18	0·04	−0·08
	3	0·41	−0·86	0·10	−0·16	−0·17	0·18
	5	−0·07	−0·06	−0·70	−0·12	0·70	−0·01
From 50 women	6	−0·09	0·25	−0·37	−0·81	−0·37	0·03
	4	0·31	−0·58	0·39	0·60	0·03	0·22
	2	0·27	0·57	−0·21	0·31	−0·25	0·63

Unfortunately, there is no test of statistical significance for Ahmavaara's method. In an attempt to determine how good is the correspondence between the rotated factors obtained from the data of the two sexes it was decided to compare the results from two successive samples of women, the first 50 and the second 70. This is found in Table 3 and, mentioning the factors from the sample of Table 3 the 50 women first each time, there is reasonably good correspondence between the factors 1 in the two groups. Factor 3 corresponds reasonably well with Factor 2, and Factor 5 is almost exactly half-way between 3 and 5. Factor 6 has reasonable correspondence with 4 and Factor 2 with 6, but has considerable weighting from Factor 2 of the group of 70. This leaves Factors 4 and 5 in the two groups unmatched. Factor 4 corresponds most closely with 4 and 2. There is therefore some uncertainty in the identification of the two sets of factors. In fact, the rotation has not improved on the resemblances found in the two sets of principal components. It would seem, on the evidence, that the Varimax method of rotation cannot guarantee the appearance of invariant factors. Perhaps the restriction of orthogonality is the difficulty, but when this is removed, other difficulties appear (Warburton, 1963). Suffice it to say that the possibility that there is an essential difference in the symptomatology of depressive states in women as compared with men cannot be excluded, although it is reasonable to say that any difference will not be great.

ACKNOWLEDGEMENTS

The major part of this work, including the computing, was done when the author was a member of the External Scientific Staff of the Medical Research Council, to whom thanks are due. The computing was done by using the P.L.U.S. system of programs. Thanks are due to members of the Department of Psychiatry for their invaluable help in discussing the work.

REFERENCES

AHMAVAARA, Y. (1954). Transformation analysis of factorial data. *Ann. Acad. Scie. Fenn.*, Series B, **88** (2), 54–59.

FRUCHTER, B. and JENNINGS, E. (1962). Factor Analysis. In H. BORKO (Ed.), *Computer Applications in the Behavioural Sciences*. London: Prentice-Hall.

HAMILTON, M. (1967). Development of a rating scale for primary depressive illness. *Br. J. soc. clin. Psychol.* (In press).

HARMAN, H. (1960). *Modern Factor Analysis*. Chicago: Univ. Chicago Press.

KAISER, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, **23**, 187–200.

KAISER, H. F. (1960). Comments on communalities and the number of factors. Paper read at an informal conference on " The Communality Problem in Factor Analysis " held at Washington Univ., St. Louis.

LAWLEY, D. N. and MAXWELL, A. E. (1964). Factor transformation methods. *Br. J. statist. Psychol.* **17**, 97–103.

WARBURTON, F. N. (1963). Analytic methods of factor rotation. *Br. J. statist. Psychol.* **16**, 165–174.

BOOK REVIEWS

Survey Sampling. By LESLIE KISH. London and New York: Wiley. 1965. Pp. xvi + 643. 83*s*.

For the non-specialist wishing to acquire the fundamentals of the theory and practice of sampling human populations, this book is very nearly the ideal starting-point. Just under half of it is concerned with the fundamental methods of survey sampling, using mathematics no more complicated than sixth-form algebra. The second half is a detailed exposition of more specialized techniques, including a long account of the practice of area sampling in the U.S.A.

In its contents, and in its level of exposition it must be compared with the two leading works on sample survey theory, Yates' *Sampling Methods for Censuses and Surveys* and Cochran's *Sampling Techniques*. The former of these is a detailed manual for survey practitioners, with little or no mathematical derivation; the latter is a thorough introduction to the statistical theory of surveys. Both Yates and Cochran are specialist works, in different senses; the one is for the specialist in practice and the other for the beginning specialist in theory. Professor Kish's new book is a nice mixture of the two, with enough theory for understanding and enough detail for practice. It has clearly been taught and re-taught to the great benefit of the text, which is dense with lively and occasionally provocative remarks. It is full of good examples and problems, and its pervading air of realism stems from the practical experience of the author. Like Cochran, Kish is concerned primarily to teach; like Yates, he provides a reference book for techniques.

This is not a book for survey theorists, as is evidenced by a sentence on page 229 which dismisses a great part of their last decade's research as " too complicated, I feel, for our purposes ". The emphasis throughout is on easily computable and reasonably efficient procedures. There are very few errors in the text (all of those which I found are corrected in an advertisement in the June 1966 number of the *Journal of the American Statistical Association*, three pages after page 575.) I should be surprised if this book did not have a very long life.

ALAN STUART.

Contrary Imaginations. By LIAM HUDSON. London: Methuen, 1966. Pp. 181. 25*s*.

Contrary Imaginations is written in a simple, lucid style, and provides a welcome contrast to much present-day writing in psychology. In this and other respects the author is an admirable exponent of his own principles. These are set out in the first chapter, and he sticks to them throughout the book—no jargon, no complex statistics remote from human subject matter, no ' blind ' theories based solely on statistics, no ' unsinkable ' or untestable theories and, on the positive side, a desire to reconcile to some extent the disciplines of mental testing and psychoanalysis. The test of these principles, of course, is where they lead him, and they lead him quite a long way. But it is, perhaps, the " no complex statistics . . . "—surely as dogmatic in its way as Watson's assertion of behaviourism which Dr. Hudson criticizes on the grounds of its rigidity—that lands him now and again in a muddle. Not only does he, perhaps reasonably for his own purposes, eschew multivariate methods of analysis (presumably because he thinks the results cannot be referred back to human subject matter) but he fails, I think, to understand the logic underlying simpler statistical methods, e.g., the effect of selection on correlation.

The main part of the book consists of a description of the results of " reasonably clever boys in the upper forms of public grammar schools " in tests of intelligence and creativity (largely borrowed from Getzel and Jackson) and in a " personal qualities questionnaire ". Dr. Hudson selected boys from the upper levels of general ability and, probably because of this, two other well-recognized dimensions of the group factor variety stand out—namely, a bias towards either verbal or non-verbal success in the intelligence test, and a bias towards divergent or convergent thinking in the tests of creativity.

Statistical tables at the end of the book give indications of these trends, and the tests (including creativity items) which demonstrate them; and for sub-groups of boys there are (i) intelligence scores of future scholars, exhibitioners and commoners at Oxford and Cambridge and of boys accepted at other and at no Universities, and (ii) sixth-form subjects chosen by convergers and divergers.

The book ends with a tentative explanation of the differences between the intelligent convergers and divergers in terms of parents' (I nearly wrote ' parental ', reading this book has cured me of that) personalities and their influence on offspring—the ' double-bind '.

Dr. Hudson's main contributions would seem to lie first in approaching his subject-matter from an essentially individual angle. This emphasizes, by descriptions of selected cases, the boys who are not neatly classifiable by the tests, and whose choice of university subject and whose class of degree may be unpredictable. Second, and following from this, there is his openness of mind in allowing, in a matter of fact way, a place for home atmosphere, emotional disturbance, and psychoanalytic mechanisms, in his individual studies of cognitive processes.

However, unless he is prepared to publish the kind of statistical tables which at least allow the reader to count heads, it is not clear how he can assess the scope and importance of the problems that he hints at in his selected case studies. As things are, it is impossible to separate his evidence from his speculations.

Nevertheless, the book is interesting, and the approach an original one. Psychologists should remember though, that Dr. Hudson is not the first to combine a genuine interest in children with the study of cognitive ability.

CHARLOTTE BANKS.

Signal Detection Theory and Psychophysics. By D. M. GREEN and J. A. SWETS. London and New York: Wiley. 1966. Pp. 455. 104*s*.

To the reviewer's knowledge, this is the only psychology book in existence dealing exclusively with the theory of signal detectability (TSD). The authors have condensed a considerable body of mathematical and experimental literature into a readable exposition of the formal TSD model and its application to psychophysical experiments. The treatment is more thorough than that given in any of the particular review articles and scattered chapters which have dealt with these problems, and the logical continuity possible in a volume of this size leaves one with the impression that in this case the whole is greater than the sum of the parts. In a word, the work is excellent. It is a milestone in the ten-year-old process of translating the mathematical and engineering concepts of the original statement of the theory into useful psychological counterparts.

Green and Swets propose that this volume can serve both as a source book and as a textbook. To further the latter purpose, they have included exercises at the end of the chapters, but since few courses are taught on the subject (and the book is expensive) it is likely that the primary impact of the book will result from its usefulness as a reference work for experimenters using TSD methodology. The present need for a comprehensive treatment of TSD, readable by psychologists, is pointed up by the number of experimental papers which devote considerable space to a review of the theory. The redundancy in such papers and also the confusing differences in measurement and notation used by various researchers could now be alleviated by the existence of this volume. Although the book does not seek to settle existing measurement problems by dictating a standard procedure, a variety of currently popular alternatives are discussed with the conclusion that more than one index of performance should be cited by experimenters. This practice should facilitate comparison of results even though individual purposes and preferences exist.

The body of the book is composed of three parts. These deal respectively with decision processes in detection, sensory processes in detection, and miscellaneous applications of TSD in psychology. Part I is largely statistical, describing statistical decision theory as a model of choice behaviour in the detection task. Topics systematically covered include alternative assumptions and their consequences, possible experimental paradigms and the form of the data expected by the theory, methods of analysing data, and comparisons among various models of behaviour in a detection experiment. This section, together with an admirably constructed appendix on experimental techniques, is sufficient to give the reader a working knowledge of the theory and its application to psychophysical problems. Although the theoretical exposition is well done, it is limited in scope, covering only the statistical decision procedures for the simple detection task and the m-alternative forced-choice task. The authors have not included any formal treatment of sequential decision procedures, an area of interest in statistical decision theory and probably of importance to psychology. The relationship of TSD to Thurstonian scaling is mentioned but not elaborated, and no effort is made to connect TSD with similar approaches to the same basic problem such as information theory and probit analysis.

As a first approximation to a description of the scope of this book, one can fairly state that few excursions are made beyond the topics contained in the collection of papers edited by Swets (*Signal Detection and Recognition by Human Observers*, New York: Wiley, 1964). As a result, every piece of formal or mathematical exposition on TSD can be (and is) accompanied by data, giving the work an apologetic air. On the other hand, such conservatism in " sticking to the data " is refreshing in view of the current rambling literature of mathematical models with more parameters than data.

The second part, dealing with sensory processes in detection, tends to be more mathematical than any other part of the book. Contrasted with Part I, which reads rather like a book on psychological statistics or measurement theory, Part II draws heavily on results obtained in electrical engineering and will probably discourage most readers without a mathematical or engineering background. An appendix is included covering enough waveform analysis to make the section understandable to unsophisticated but determined readers. Part II is a graphic reminder that TSD is a stimulus-oriented theory requiring a description of the signal and its noise background and leading to a specification of the performance by an ideal observer. This approach means that there are very few free or hypothetical parameters, since most parameters can be measured at the input to the observer. The freedom allowed by the theory is in the form of various levels of uncertainty about the signal and internal noise which can be postulated to degrade optimal performance in such a way that it describes human performance. The price of such a theory in psychoacoustics is apparently some complicated mathematics. Ideal performance is more difficult to specify for the visual modality, and there has been hardly any analysis of this kind for other modalities. The approach using the ' ideal observer ' concept is one of the singular contributions of TSD to psychology, and this section, while dealing with a limited scope of substantive problems in psychoacoustics, indicates the potential of this type of theory.

Part III is a wide ranging discussion of problems in perception covering multiple observations, frequency analysis, and speech communication in greatest detail while touching on many others. Here again, the authors appear to have chosen much the same topics as were contained in Swets' collection of readings, but a considerable quantity of new data is included. By the reviewer's rough count, half the graphs of data in the section are also contained in the Swets' collection.

The obvious effort which was expended in making this volume complete in itself tends to underline the few omissions one finds. In the reviewer's estimation, the most serious omission is Elliot's tables of d' (Swets, 1964). Green and Swets do discuss a graphical method of obtaining d' for the simple detection task which is also applicable to the 2-alternative forced-choice task, but for tasks involving a higher number of alternatives the reader is referred to Elliot's tables. In view of the wide usage of the m-alternative forced-choice paradigm for the study of short-term memory and the effects of uncertainty on detection, this omission is rather unfortunate.

PATRICK A. LUCAS.

S.P.

Logical Foundations of Psychological Measurement. By STRANGE ROSS. Copenhagen: Munksgaard. 1964. Pp. 152. 36 kr.

A number of significant contributions to psychophysics and measurement theory have come from Scandinavia in the last decade. The majority are from Sweden and are consistent with, or at least sympathetic to, S. S. Stevens; this monograph is from Copenhagen and follows the more traditional position of Campbell.

The author moves in four steps towards what he considers to be a demonstration of the inadequacy of Stevens' position; a review of semantic systems and propositional calculi leading to formal definitions of representation and isomorphism comes first and follows Carnap or Tarski on many points. This review should be useful in itself as a background reading for mathematical psychology courses if students have already done about a year's symbolic logic and have used a notation similar to that in Copi's standard text.

From his axioms Ross then outlines elements of a philosophy of physical measurements. Ordinal and additive measures are separately treated in that order, each being introduced through appropriate calculi and interpretations of these. As might be expected from a follower of Campbell, additivity plays a central part in Ross's measure theory.

In his calculi of addition, Ross introduces the ' axiom of continuity ' and notes that it creates difficulties; he uses weight as a topic for a measurement theory in physics and immediately he is obliged to introduce the possibility in statement $S_2 4$ (page 53) of breaking up given objects, an operation which may not have an analogue in psychological measurement. This would not matter, except that in chapter three Ross tries to parallel his treatment of physical measurement in his psychological theory, and his treatment centres on Campbell's distinction between extensive or fundamentally measurable dimensions and derivative measures.

His determination to subsume physical and psychological measurement under similar formal structures is nicely put (page 68), where he writes " we will assume that all knowledge within any branch of science may be stated by axiomatized theories of the structure previously described *in this monograph* (my italics) . . . consisting of a semantic system and a calculus (and a set of constancy rules). As a calculus is completely independent of any interpretation it follows that it will not be possible to discriminate between the different branches of science on the basis of the calculi employed within these branches." From this position Ross is led, willy-nilly, to define psychological measurement theory as a psychological theory whose calculus is isomorphic to either the calculus of weak order or one of the calculi of addition. Other calculi which have been considered by Suppes and the scale types described by Coombs, which should have their own unique calculi, are passed over. The admission of their existence is presumably damaging to Ross's thesis of isomorphism between physical and psychological measurement.

Ross is quite willing to allow the primitives of his calculi to be given non-physical interpretation in a psychological theory, and consequently runs directly into the problem of unreliable or unreproducible data which he circumvents with the introduction of three stochastic calculi which serve for ordinal measurement. Three theories, each of some complexity at the axiomatic level, are required; the first to achieve a partial invariance over stimuli, the second to do the same for responses and the third to combine the two, so that its set of sufficient conditions for falsification is the union of the sets for the preceding two. This enables Ross to cope with ordinal measurement at the level of, say, an intelligence test ordering of subjects. He then states and offers a proof of a theorem in the general theory of measurement: " if two or more measurement theories of the additive type may be constructed on the basis of a given . . . theory of the ordinal type, then the *same* members will be assigned the *same* elements through *all* these measurement theories of the additive type . . . on the basis of a given measurement theory of the ordinal type (at most) *one* additive scale may be found ".

In axioms of the general theory, Ross gives (page 110)

$$\text{NAA3} \quad N(z) = N(x) + N(y) \equiv zP(xAy)$$
$$\text{NAA*3} \quad N^*(z) = N^*(x) + N^*(y) \equiv zP(xA^*y)$$

where A and A^* are operations of combination N and N^* are number assignments according to the two methods involved, and P is an equivalence operation corresponding to ' = ' in the number system. He then takes two elements s_0 and $s_0{}^*$ such that $s_0Ps_0{}^*$, and applying NAA3 gets $0_4P(s_0As_0{}^*)$, from which he then gets

$$(3.10.6) \qquad N\ (s_0)+N\ (s_0{}^*)=N\ (0_4)$$

and

$$(3.10.7) \qquad N^*(s_0)+N^*(s_0{}^*)=N^*(0_4).$$

However, as Eisler in 1965 pointed out, (3.10.7) implies $0_4P(s_0A^*s_0{}^*)$ and we cannot properly state that $0_4P(s_0As_0{}^*)0_4P\supset(s_0A^*s_0{}^*)$. On the contrary, it is empirically very unlikely that two measurements systems differing in A and A^* will yield the identity $N(0_4)=N^*(0_4)$.

So the central theorem on which Ross's thesis of an isomorphism between physical and psychological measurement rests is invalid. Ross's reductionist attempt at theory leads him to state, in effect, that if a psychological scale of x-ishness exists it would necessarily be the same as the physical scale of x-ishness. Obviously this position is in head-on conflict whit Stevens' rationale for constructing psychophysical power functions, and Ross is concerned to offer an account of what Stevens is actually doing when he uses magnitude estimation or production. Ross does this by asserting that Stevens is developing a subjective and not an objective psychology: " the criticism of the conception (presumably) underlying Stevens' scaling procedures may be directed at the most extraordinary fact that according to the conception, it is left to the subject to perform those operations performed exclusively by the experimenter within an objective psychology ". The subject's statements (about stimulus magnitudes) no longer belong to the object language, but are on a par with the statements of another psychologist and thus belong to the metalanguage.

Ross ends up, therefore, in a logical position which he could have arrived at without his prior excursion into measurement theory, a position in which his definitions of the logical status of subjective experience or the reports of such experiences in psychology necessarily determine what shall and what shall not be measurement. This book is, in a way, five years too late; whilst Ross has been stating what is measurement, Suppes and Zinnes have made contributions which strongly suggest that the insistence on additivity in Campbell's view of measurement is not well put, and independently Goude has, in any case, found an empirical operation corresponding to numerical addition that can be made within the context of Stevens' psychophysical procedures. It is sad to reflect that if F. P. Ramsay had not died early, before he had finished the demolition of Campbell, which began in a little known paper of his on Chance, then we might have been spared all this effort and confusion.

R. A. M. GREGSON.

Pattern Recognition: Theory, Experiment, Computer Simulations and Dynamic Models of Form Perception and Discovery.

Edited by LEONARD UHR. London and New York: Wiley. 1966. Pp. ix + 393. 68*s*. (cloth), 45*s*. (paper).

This book on the recognition process in man and machine is a collection of previously published papers. But unlike most such collections, where the papers report the use of similar methods to investigate closely related questions, it provides a selection of the material with which a research worker in this relatively new interdisciplinary field ought to be familiar. The selection comes from philosophical, psychological, physiological and engineering books and journals.

The book consists of five sections—conceptual framework (mainly philosophy), empirical background (psychology), theoretical developments (mainly speculative physiological psychology), experimental results pertinent to model building (neurophysiology) and computer simulation; each section contains between four and six papers of a high quality. Unfortunately only a few of these papers are surveys of any sort. There is a very competent survey of computer simulation by Uhr himself, and the section concerned with the empirical background to the study of recognition contains four papers of the 'Psychological Bulletin' variety by Vernon, Wohlwill, Attneave and Arnoult, and Hake. The paper by Hake on form discrimination and the invariance of form is especially valuable as it is a very thoughtful survey which has been published previously only as a technical report.

Many of the other papers are important contributions to one of the disciplines which are involved in the study of pattern recognition and were therefore written for the appropriate specialists. However, as the area is so wide they provide only a very small selection of the important papers that could be covered, so the expert will probably not make great use of the book. Unfortunately, to the uninitiated they may well prove both difficult and boring because they are not edited; and not only are some of them highly technical but they are also relatively unconnected, which makes interest much harder to sustain. This might have been remedied by greatly lengthening the introductory notes to each section, three of which are only just over one page long.

This reviewer considers that it would have been better if a single author had been persuaded to write a book covering the present topic; but this volume does serve a valuable function. It delimits a relatively unstructured interdisciplinary field and focuses more attention on a very important and almost totally unanswered problem. It is probably a consequence of it being an edited selection of papers that one area which might well provide very useful lines of attack was not mentioned. This is the study of the mathematics of pattern recognition; e.g. the work of Nilsson. Yet to have added mathematical papers to the others would possibly have been the last straw for many readers. T. SHALLICE.

Contributions to Mathematical Psychology. Edited by NORMAN FREDERIKSEN and HAROLD GULLIKSEN. New York: Holt, Rinehart & Winston. 1964. Pp. x + 189. 54*s*.

This is a symposium of papers in honour of L. L. Thurstone presented when Thurstone Hall was dedicated to him by the Educational Testing Service. The quality of the papers, the academic distinction of the authors (who were his students or strongly influenced by him) and the different universities they represent form a magnificent tribute to Thurstone himself and the spread of his influence.

Five of the papers are mainly mathematical in their content, but two, the first and the last, are not. The first, by Dorothy Atkins, North Carolina, is an account of Thurstone's professional career, with a few sidelights on his private life. Appended to it is an exhaustive bibliography of his publications, which requires sixteen pages to record his huge output of work.

Thurstone was born in Chicago in 1887. His parents, who had come from Sweden, took him back to Stockholm when he was eight but returned to the U.S.A. with him and settled in New York when he was fourteen. At the age of twenty-one he entered Cornell University, to take courses in civil engineering and later in electrical engineering. While there, he designed a motion picture camera and projector. He succeeded in arranging to demonstrate it to Edison and his staff; and Edison offered him an assistantship in his laboratory, though he did not use the design. Thurstone showed an early interest in the psychological aspects of his work, but it was not till 1914, when he was twenty-seven, that he began to concentrate on psychology. From the first, he was interested in the mathematical approach to the subject, and his thesis was about the fitting of hyperbolae to records of typewriter learning. Throughout his career he continued to make vigorous use of mathematics for investigating the problems which interested him in other fields, but he does not appear to have specialized in mathematics at any stage or to have considered its implications as much as its applications. He was introduced to matrix algebra through a chance conversation with colleagues after his appointment as Professor of Psychology at Chicago and not long before he published *The Vectors of the Mind* in 1935.

An interesting feature of most of the other papers in this symposium is that they are more in the nature of progress reports on the work of the authors than final reports of limited research projects. So they provide excellent, brief and authoritative outlines of the authors' approaches to their various subjects.

Coombs, Michigan, summarizes his theory of data before going on to a non-metric multidimensional analysis of a conditional proximity matrix. The example he gives is a non-symmetrical square matrix in which each row and corresponding column refers to an American psychological journal, and the characteristic entry in cell i, j is the frequency of references to articles in journal j appearing in articles in journal i. He found the data fitted the requirements for a metric space well, with one dimension predominating, which runs from the *Journal of Abnormal and Social Psychology and the Journal of Consulting Psychology* at one end to the *Journal of Experimental Psychology* and the *Journal of Comparative and Physiological Psychology* at the other.

Gulliksen, Princeton, gives a discussion of some of the problems encountered in analysing statements of personal preferences and follows it with an analysis of a massive collection of data from groups of students in Belgium, France, Germany, Italy and Texas, U.S.A., who were given four questionnaires concerned respectively with the evaluation of occupations, reasons for work, nationalities and goals in life. The obtained values were scaled and their means calculated for each group. Some comparisons between the groups are recorded graphically. The data for Belgium and Texas were used for two factor analyses of the intercorrelations between the items in all four questionnaires combined. A rather striking finding was that there is little tendency for the rotated factors to have loadings on more than one of the questionnaires. Thus, for example, evaluations of the goals of serving God and gaining immortality, which characterize factor 6 in the Texas data are reported to be unrelated to evaluations of the occupations of missionary and clergyman which characterise factor 3.

Tucker, Illinois, tackles the problem of extending factor analysis to three-way tensors (which he calls, rather confusingly, three-dimensional matrices) and provides a method of reducing a table of the measurements of a number of variables in a number of cases on a number of occasions to a smaller three-way table, or core matrix, which may perhaps be capable of being interpreted in the usual factor-analytic terminology. But as his example is a fictitious one the psychological relevance of his results is hard to judge.

The problem discussed by Horst, Washington, is one met in applying factor analysis to the items in a test or questionnaire scored dichotomously. It was Guttman who first examined the implications of the concept that items must belong on a single scale if there is any way of arraying them in order of bias, e.g. from easiest to hardest or from least favourable to most, such that whoever gives a positive (correct or favourable) answer to an item anywhere in the order has given a similar answer to every previous one: that is to say, if the earlier of any pair of items receives $m+n$ positive answers and the later one n only, they must belong to the same scale if there are m informants who give positive answers to the earlier one but not the later, and none who give positive answers to the later but not the earlier. Although this simplex phenomenon is obviously valid evidence that the two items can be referred to a single scale, a four-point or product-moment correlation between them equals unity only when $m=0$, and diminishes steadily as m increases. As Horst points out, the rank of the matrix of correlations between the items equals the number of different divisions they mark on the scale, and reduces to unity only in the trivial case where they all divide it at the same point and the range of scores over the test as a whole becomes limited to two values. Item analysis thus reveals a conflict between Guttman's definition of a single scale in terms of the simplex phenomenon and the factor-analytic concept of a single dimension of variation. The conflict can be avoided by using tetrachoric instead of four-point correlations; but this amounts to suppressing the evidence of variation in bias though " we know that it is an essential part of the warp and woof of all tests made up of dichotomously scored items ". When the items may refer to more than one scale and the status of some of the items in relation to some of the scales is open to doubt, that is to say in the typical conditions where a multifactorial item-analysis is needed, the problem becomes more complicated. Effects of differences in bias need to be disentangled from effects of differences in scales. Horst proposes a procedure to partial them out and obtain a residual matrix for factoring.

Abelson, Yale, discusses changes in the distribution of attitudes that may occur when a controversial issue is discussed between members of a social group. He makes use of contact matrices, which are square matrices, say A, with a row and corresponding column for each member of the group, and are symmetrical if the entry in cell i, j and again in cell j, i is the frequency of contact between members i and j. These off-diagonal entries, for $i\,j$, may be positive or zero but cannot be negative; the entries in the leading diagonal, for $i=j$, may be taken to be zero if it is considered reasonable to assume that contacts of i with himself do not produce any change in his attitude. The initial attitudes of i and j towards the issue are assumed to be measurable as x_i and x_j, and the change in the attitude of i as a result of contact with j over time t is expressed by the differential equation

$$\frac{d_j\, x_i}{d_t} = ka_{ij}\,(x_j - x_i),$$

where a_{ij} is the entry in cell i, j.

This leads Abelson to consider A^*, a derivative of A, with the same off-diagonal entries but with negative entries in the leading diagonal which reduce the row totals to zero. He analyses it into components with real bipolar vectors and negative roots. The vectors define the divisions in the group: contact is relatively close between members located at the same pole and slight between members at opposite poles; the first component (which has the smallest negative root) gives the major cleavage, etc. The initial attitudes of the members towards an issue will be the slower to change, the more closely the vector of their measurements coincides with the major dimension of cleavage. The model implies that all the values of x converge asymptotically, on a mean i.e. that the members all ultimately reach agreement, provided there is at least one who is in contact, directly or indirectly, with all the others. Finally, Abelson proposes extending the model by introducing many further assumptions.

The last paper, a literary one by Guilford, Southern California, is on the nature of creative processes. He relates them to his general scheme for analysing intellectual processes, viz. into four kinds according to their content, five according to the operations they involve, and six according to their products—a scheme which leads him to expect that factor-analysis will ultimately identify up to $4 \times 5 \times 6 = 120$ primary intellectual abilities. Some 60, already identified, fit into the scheme, and it has proved its heuristic value as a source of hypotheses concerning areas to investigate in search of other so far unidentified ones. Creative processes, he considers, result from combinations of these abilities, some contributing more than others to the outcome. The abilities involved in divergent production (generating a variety of additional items of information from given ones) are particularly important. Those involved in the evaluation of implications and semantic systems are also needed. So, too, are those in the convergent production of transformations (notably redefining miscellaneous items to generate a new fully-determined one). Guilford accepts Graham Wallas's account of different stages in the creative process—preparation, incubation, illumination and verification—with some slight reservations, and discusses which of the abilities are the most important ones at each stage. From this standpoint he comments on Spearman's and Thurstone's theories of creativity, as well as psychoanalytic theories. He concludes by indicating some of the problems needing investigation in this field.

The papers cover such a diversity of subjects that there must be many readers who will need to consult or study some of them, and the book is likely to remain important as a reference for a long time. The papers on subjects which are unfamiliar to me are the ones I find most interesting; the ones on subjects I have studied closely are relatively disappointing.

Abelson's differential contact frequency matrix fascinates me with its very peculiar properties and I would have welcomed more information about his methods for analysing it. Coombs's conditional proximity matrix interests me only slightly less. The main attraction of both their papers is the prospect they offer of widening the scope of scientific observation. They provide methodological scaffolding from which new vistas for research can be seen. It would not matter much to me if the scaffolding had some day to be pulled down to make place for more solid structures built on firmer foundations. The possibilities they reveal are their attraction.

On the other hand, when considering the papers by Gulliksen, Horst and Tucker, I find myself more critical of the structures, not being so preoccupied with the view. Their methods seem quite unnecessarily elaborate and not very appropriate for their data. Gulliksen's materials are so vast that his paper, long as it is, does not explain fully how he analysed his data or show what they were before he began. I can see no advantages whatever in the equamax rotation he used for his factor-analyses: product matrices such as his must be singular, provided the effects of response set are removed, so the communality problem does not arise and his procedure starts with a false assumption; moreover, a rotation which tends towards a positive manifold (there are less than 10 negative loadings in the 300 or so listed in his Tables 11 to 14) is inappropriate for expressions of preference, since they are necessarily bi-polar. I have described elsewhere the methods which I prefer (Slater, 1958, 1960).

Horst's problem appears to me to be one that might be solved by arraying the observations in a matrix with a row for each informant and a column for each item, using analysis of variance to eliminate the variation between rows and columns, forming a product matrix from the interactions and analysing that in terms of its principal components (it, too, is a singular matrix, and any analysis introducing reduced communalities would be inappropriate).

For three-way tensors a wide range of possible methods of analysis are worth considering. Several are being used in the M.R.C. service for analysing repertory grids. The most promising one, however, and the one which is most relevant to the problem discussed by Tucker, is still undergoing development. It depends on finding a convenient working method for solving the equation

$$T(o, n, m) = \lambda^{\frac{1}{2}} . v_o . v_n . v_m + E(o, n, m).$$

Here T is a three-way tensor built up of a series of o matrices with n rows and m columns each, e.g. a set of repertory grids obtained from the same informant, using the same constructs and elements on o occasions during the course of a psychiatric treatment. It is referred to a principal component specified by three vectors, namely v_o for the series, v_n for the rows and v_m for the columns, and a latent root, λ. E is a three-way tensor of residuals; and the specifications for the component must be chosen to minimize the sum of the squares of the entries in E. Dr. J. C. Gower, of the Department of Statistics, Rothamsted Experimental Station, found a feasible algebraic solution for the equation when I proposed the problem to him in the spring of 1966, and his paper on it is due to appear in *Biometrics*.

This is the orderly way of extending principal component analysis to three-way tensors; and it can be carried further to four-way and higher order tensors by following the same rule. The problem is far from being academic: for example, a set of p three-way tensors, obtained from p patients undergoing a course of group psychotherapy, may be defined as a tensor with a fourth subscript, p, and referred to a component specified by four vectors, including v_p for the patients, plus a four-way tensor of residuals with a minimized sum of squares. Singularities present no difficulties, which is fortunate since the higher the order of the tensor the more singularities must occur.

References

GOWER, J. C. Generalizations of principal components analysis. *Biometrics.* (In Press).

SLATER, PATRICK (1958). The general relationship between test factors and person factors; application to preference matrices. *Nature*, **181**, 1225–1226.

SLATER, PATRICK (1960). The analysis of personal preferences. *Br. J. Statist. Psychol.* **13**, 119–135.

PATRICK SLATER.

Multivariate Statistical Analysis for Biologists. By HILARY E. SEAL. London: Methuen. 1964. Pp. xiv + 207. 45*s*.

This book is written for the biological research worker who has little understanding of least squares and multivariate models. It has the aim of affording him " a sound working knowledge of . . . Regression Analysis of a Linear Model " and of explaining " multivariate statistical methods in relatively simple language which yet omits none of the technical details ". The role of the computer in statistical analysis is at once acknowledged and it is rightly pointed out that there is danger in the ready availability of programs for multivariate analyses, which may be used indiscriminately with only a dim notion of the underlying model.

There are two sections, Part A entitled " A Single Dependent Variate " and Part B entitled " Several Dependent Variates ". In the first chapter the reader is at once confronted by a set of data for which a linear model is developed, and the remainder of the chapter takes him through the analysis of this data in a conversational tone which characterizes the entire volume. The topics of least square estimation, vector notation, vector addition and multiplication, solution of simultaneous linear equations, determinants, and estimation and distribution of residuals, constrasts, and regression coefficients are treated within the first 25 pages. The second chapter defines orthogonality of vectors and derives four linear models for a further set of data. It is shown that orthogonality of the instrumental variates simplifies computation, which seems a slight virtue for computer users, but it is not mentioned that this minimizes sampling variance of the regression coefficients. The third chapter deals with analysis of covariance, and the fourth, misnamed " Multiple Regression ", really deals with matrix algebra up to matrix inversion.

Part B, on multivariate analysis, commences with a chapter on multivariate analysis of variance. The following chapter outlines principal components and introduces eigenvalues and eigenvectors and the multivariate normal distribution. The next chapter, entitled " Canonical Analysis ", covers what is usually referred to as discriminant analysis and has nothing to do with canonical correlation. A test for the equality of covariance matrices is introduced here. The final chapter is a muddled treatment of factor analysis and will disappoint most psychologists familiar with this subject.

Because of the shortness of this book, it is well to consider what is omitted, especially with respect to Part B. There is no mention of the test of the difference of a mean vector from zero or from another mean vector with or without a known covariance matrix. The Hotelling T^2 statistic seems more easily treated and more applicable than, say, discriminant analysis. The multivariate Behrens–Fisher problem and a test for proportionality of one covariance matrix to another would also have been very useful. The omission of canonical correlation has been already mentioned. Within the context of discriminant analysis, further consideration of the classification problem might have been desirable. Multiple regression was also rather slightly treated and partial correlation is not mentioned. The book as a whole lacks any statistical tables or a list of references for the over-all subject. The reader is nowhere told of a more complete treatment of either least squares theory or multivariate statistics. For example, there is no reference to Anderson's 1958 textbook.

This work contains several blunders in numerical analysis. The solution of three simultaneous linear equations by Cramer's Rule is described for " the insight it gives us into electronic machine computations for solving such equations ". The solution for eigenvalues by finding the roots of the characteristic polynomial is outlined and only in the appendix is the reader referred to the Jacobi method by way of a paper by Ralston and Wilf, published in 1960. (In fact, many computer centres will now insist on the much more efficient Householder-Ortega-Wilkinson method described by Ralston in 1965.) In the factor analysis chapter it is not mentioned that the iterative solution for communalities, besides being very time-consuming, will occasionally yield values greater than unity and that the degree of optimality of these estimates is unknown.

It seems unlikely that the reader with only " a one semester course in basic statistics " will gain more than a superficial understanding of the models described. The author's breezy informality could be deceptive and there is a danger that the indiscriminate use of multivariate analysis programs will be encouraged rather than discouraged by reading the book. There is very little discussion of when models and tests are inappropriate due to, say, violations of independence or sub-asymptotic sample sizes. On page 115 the asymptotic distribution of the likelihood ratio criterion is employed with an N of 24.

If the aims of the book had been fulfilled, it would have been a valuable aid for psychological research workers. But they are not; and the book contains some bad statistics and numerical analysis. It does, however, have many sets of illustrative data—these are actually its only biological features. It may therefore be useful as supplementary material for a more complete text.

J. O. RAMSAY.

Introduction to Statistical Ideas for Social Scientists. By G. KALTON. London: Chapman & Hall. 1966. Pp. 58. 6*s*.

This is a short primer in which the author has attempted to provide a broad review of statistical ideas. His explicit aim was to provide an introduction to statistical thinking without the complication of detailed computations. The reader will probably acquire some knowledge of basic statistical notions and will learn to compute basic descriptive statistics. Statistical testing procedures, including chi-square, Students *t*, product-moment correlation and the logic underlying simple one variable analysis of variance, are introduced. However, it is not the author's aim that this book alone should prepare an otherwise uninitiated student to use these procedures. There are several references to more advanced texts. No statistical tables are provided.

It is this reviewer's opinion that if a student is to learn statistical testing procedures competently, he will need access to more advanced texts than this one, and that many such texts provide as good a grounding in fundamental concepts as does this book. The aim to provide a clearly stated introduction to statistical ideas is praiseworthy, but would be more likely to succeed if greater space had been devoted to it. If supported by an appropriate lecture course, this book may be very valuable to the social science students for whom it was written. However, it is quite possible that the main value of this book will be to students who have already learnt some simple procedures but who have continuing difficulty in understanding what is accomplished by their use.

D. LEGGE.

On Theory and Verification in Sociology. By H. L. ZETTERBERG. 3rd Edition. New Jersey: Bedminster. 1965. Pp. 177. 40*s*.

This is an elementary and highly readable introduction to theory building in the social sciences. For the third edition a new chapter on *definitions* has been added, and the text now touches lightly on a whole range of theoretical structures—from simple inventories of findings to complex axiomatic systems and computer simulations. In addition, the relation of theories to facts is discussed.

Most of the book is concerned with simple ' covering law ' types of exploration, in which qualitative statements are subsumed under, or deduced from, other qualitative statements. No attempt is made to go beyond existing textbooks, or to criticize the adequacy of the kinds of theorizing they describe. But the book makes some telling points, particularly against the kind of textbook which contains a mass of definitions and propositions that are never integrated with one another. It is well supported with illustrative materials, and it conveys an accurate image of the kind of 'propositional sociology' favoured by an increasingly influential group of American sociologists. B. N. LEWIS.

Memory Abilities: a Factor Analysis. By H. PAUL KELLEY. Psychometric Society: Monograph Number II. 1964. Pp. xi + 53. $3.00.

This monograph contains a report of a factor analytic study of what the author describes as "relatively immediate intentional retention" as distinct from 'incidental or long-term memory'. The purpose of the study is to find out whether *rote memory*, *meaningful memory*, *span memory* and *visual memory* are distinct and separate mental factors, independent of the modality of presentation of the test material and of the type of material used.

Having defined each of these factors, the author proceeds to construct a battery of tests which should sample the factors postulated and also differentiate between them. Realizing that this battery of tests is likely to measure abilities other than memory, he supplements it with some reference tests from the U.S. Air Force Airman Classification Battery. The final battery of 40 tests was administered to a sample of 442 U.S. Air Force pilot cadets, a correlation matrix was obtained and 11 factors were extracted by Thurstonian methods. The factors were then rotated to oblique 'simple structure', partly by mechanical and partly by graphical methods. The first three factors obtained are identified as *verbal comprehension*, *numerical facility* and *perceptual speed*. Their extraction may be thought of as a purifying process, and once they are out of the way three factors which the author labels *rote memory*, *meaningful memory* and *span memory* emerge. These are three of the four memory factors postulated: they appear to be relatively distinct from each other and to be independent of whether the material is presented visually or in an auditory manner. The evidence with regard to the postulated *visual memory* factor is not found by the author to be convincing.

All this sounds fairly satisfactory and in a sense it is. Yet on reading the monograph some of the old apprehensions about the value of factor analytic studies seem to well up. Here one should note that although the monograph was published in 1964, the actual study was completed ten years earlier. As a consequence, the statistical procedures used are now somewhat out-dated. Since the study is confirmatory rather than pilot in nature, and since predefined factors have been postulated to emerge from it, a competent factor analyst today would postulate a matrix of zero and non-zero loadings for the data and proceed to test his hypothesis by appropriate methods. This would eliminate the natural doubts one has about the objectivity of graphical methods of rotation in hypothesis testing. However, having noted these reservations, it is only fair to say that this is an interesting monograph and is likely to be widely read by educationalists as well as by psychologists interested in the study of memory.

A. E. MAXWELL.

Questionnaire Design and Attitude Measurement. By A. N. OPPENHEIM. London: Heinemann. 1966. Pp. ix + 298. 35*s*.

This is essentially a 'how to do it' book likely to be of greatest use to the practitioner in the social sciences. It represents a readable account of the hazards that face the would-be designer of a questionnaire. "It is not meant to be a work of scholarship, nor is it exhaustive". The strong point of the book is its emphasis upon the wording and testing of the individual items which go to make up a questionnaire or attitude scale.

As a possible "textbook for research students and undergraduates in social psychology and sociology" it has severe limitations, particularly in the area of attitude measurement.

The discussion in Chapter Five of the concept of 'attitude' is weak; and the account in the following chapter of the relationship between attitudes and behaviour is very inadequate. Perhaps these are two areas where the writer has exercised his option " to omit some of the more esoteric controversies ". A full discussion of these two issues is vital for the student and research worker in social psychology, and the second issue is of central importance to the practitioner who is concerned with predicting behaviour from a knowledge of attitudes.

The attitude scaling procedures outlined in detail are the traditional Thurstone, Likert and Guttman scales. More recent multidimensional approaches to attitude scaling do not receive the same pride of place. The Semantic-Differential and Repertory-Grid techniques are linked together in a chapter on " Miscellaneous Techniques " which also includes sociometry, stereotypes and brand-images, and diaries. Another odd juxtaposition is the inclusion of a section on " experiments " appended to the chapter on projective techniques.

The Bogardus Social Distance Scale appears in its original 1925 version and the interested reader is referred to selected readings outlining more recent research on the topic. The Selected Readings at the end of each chapter often appear to make up for known deficiencies in the text, e.g. a single paragraph of text on response sets is supplemented by reference to six selected readings. This is particularly surprising in the light of the author's claim to have kept the number of references to a minimum.

Attitude scaling methods are approached in a refreshing way through a full discussion of the wording of questions and the writing of attitude statements. The importance of pilot work in testing the wording of individual questions and the optimal sequencing of items in the questionnaire is stressed throughout the book. The examples come from a wide variety of settings though mostly from surveys in which the author has been involved.

There is a full chapter on the use of projective techniques in attitude study. It would have been more helpful here if the author had given some guide lines for the coding of qualitative material similar to his excellent step-by-step outlines for the development of attitude scales. Admittedly this is rather a tall order, but there are a number of books and articles on content analysis which are likely to be of more use to the practitioner than the present account.

The final chapter deals with the quantification of questionnaire data. Once again the outlining of a set of standard operating procedures would have been more helpful than a lot of text supported by a full-page diagram of a blank punch card.

R. M. FARR.

Psychological Tests. By EDGAR ANSTEY. London: Nelson. 1966. Pp. 300. 63*s*.

Writing for " professional test constructors and for the many people who are concerned with intelligence and other psychological tests ", Dr. Anstey is at his best when he discusses the content of test items and the many faults that may be discovered by thoughtfully applied item analysis. Many examples are provided and their detailed examination makes a useful and perhaps unique contribution. As for the rest, a serious lack of scholarship is evident in most areas. The theoretical shortcomings are too numerous to discuss in any detail, but a few pointers will suffice.

Firstly, the title is misleading. As Dr. Anstey acknowledges in the Preface, most of the content is about item analysis and intelligence tests. Secondly, the discussion of the alternative indices for item analysis, to which a good deal of space is allotted, is in most respects out-dated. The availability of computers has removed the need for 'short-cut' methods among professional test constructors. Moreover, a comparison of the indices in terms of their mathematico-statistical properties is a pointless exercise when no explicit reference is made to a test-theoretic model.

Thirdly, there is a strange if not musty flavour to the references quoted on many topics. On item analysis we are primarily referred to Burt (up to 1945), Guilford (1936, 1941) and Vernon (1945, 1948). On corrections for guessing in multiple-choice tests the main reference is to Davis (1951). (It is not good enough to say that sufficient has been written on this topic and then to give an inadequate and even misleading summary of the problem). On reliability we are referred to Adkins (1947), Guilford (1936) and Thorndike (1951). On assessing the value of tests, the important contribution by Cronbach and Gleser (1957) is not discussed although it is listed in the Bibliography. On multiple regression we are referred to Weatherburn (1946) and an unpublished note by Maund (1951). On the test-theoretic definition of reliability, Thorndike (1951) is credited rather than Spearman, Kelley or even Gulliksen. On item analysis by computer only Ziegler (1954) is mentioned. The brief index lists Socrates, Inigo Jones, decor of office premises and similar oddities, but not Spearman, Thurstone, Gulliksen, aptitude tests and test utility. All in all, Dr. Anstey has a good eye for the practical points, but the scholarship displayed is weak.

PHILIP LEVY.

COMPUTER PROGRAM ABSTRACTS

Program for Factor Analysis. By M. WYATT, Department of Education, University of Manchester, Manchester 13.

Description This program carries out a principal components analysis, rotates the axes in accordance with the Varimax criterion and then uses the Promax method to obtain oblique simple structure. The intercorrelations of the oblique factors are analysed for higher order factors, each higher order factor being given in terms of the original variables. The starting point of the program may be either correlations or raw scores. In the latter case, means, standard deviations and correlations are computed and printed out.

The maximum number of variables acceptable to the program will be approximately 120 to 130 on the Manchester University Atlas.

Computer, Language, Data Format The program has been designed for use on the Manchester University Atlas and is written in Atlas Autocode. The program is on 7-hole paper tape. Data input may be either 7- or 5-hole paper tape or punched cards. A data-checking system is incorporated in the program for use when starting from raw scores, a subject's scores being deleted from the computation if they are found to be at fault. The nature of the fault and location are printed out.

The program will analyse a number of separate sets of data, the number being given at the commencement of the data.

Computation Times Most average size analyses are completed in one minute. For example, analysis of a data matrix of 100 subjects scored on 30 variables requires approximately 30 sec. Analysis of a data matrix of 200 subjects scored on 45 variables requires approximately 90 sec.

Availability Program tapes and directions for use are available from the author at the Department of Education, University of Manchester.

For use of the program on the Manchester University Atlas, enquiries should be addressed to Head of University Computing Service, Manchester University, Dover Street, Manchester 13.

Reference

HENDRICKSON, A. E. and WHITE, P. O. (1964). Promax: A quick method for rotation to oblique simple structure. *Br. J. statist. Psychol.* **17**, 65.

A FORTRAN Program for Canonical Correlations. By S. HUNKA and M. WAHLSTROM, University of Alberta, Edmonton, Alberta, Canada.

Description This program performs a canonical correlation analysis of two sets of variables as described by Cooley and Lohnes (1962) and Anderson (1958).

Language and Configuration The program is written in FORTRAN IV for use on the IBM 7040 having 32K core storage. No tapes are used. Variable formating is used and the total program is composed of 13 subroutines.

Timing Three problems having 170 observations, each with a total of 11 variables, required 11 sec. of computer time. This estimate includes calculation of correlation coefficients.

Input, Options, Output Input must be on cards, but may be either raw data or the appropriate correlation matrix. A data-transformation subroutine is made available for rearranging raw data to form the required sets of variables.

Various options are available for the data to be output. If raw data are used, the means, standard deviations and correlation matrix are output. Regardless of whether raw data or a correlation matrix is input, the partitioned matrices are output. In addition to the determinant of the correlation matrix being output, inverses of the partitioned matrices, and intermediate calculations required to make symmetric the matrix for which roots and vectors are required are also given. In all instances where eigenvalues and eigenvectors are found they may be output. Canonical correlations and both sets of normalized weights are output and properly labelled. Bartlett's Lambda test is used to test for the significance of the canonical correlations.

Availability Copies of the source deck or listings may be obtained by writing to the Division of Educational Research Services, Education Building, University of Alberta, Edmonton, Canada.

References

COOLEY, W. W. and LOHNES, P. R. (1962). *Multivariate Procedures for the Behavioral Sciences*. New York: Wiley.

ANDERSON, T. W. (1958). *An Introduction to Multivariate Statistical Analysis*. New York: Wiley.

A FORTRAN Program for Nonparametric Statistics. By J. E. CARLSON, University of Alberta, Edmonton, Alberta, Canada.

Description This program can be used to calculate various nonparametric statistics as described by Siegel (1956) and by the Application Directory of the IBM System/360 Scientific Subroutine Package. The following statistics can be calculated in a given run:

(1) Rank correlation coefficients between all pairs of variables (either Spearman's ρ or Kendall's τ).

(2) Any one of the following:
   - (*a*) Chi-square test between any two variables, categorized at previously specified score points.
   - (*b*) Cochran's Q-test between dichotomous variables (minimum of three variables).
   - (*c*) Mann–Whitney U-test between two specified groups on each variable.
   - (*d*) Friedman's Two-Way Analysis of Variance between matched groups.
   - (*e*) Kruskal–WallisAn alysis of Variance between specified groups on each variable.
   - (*f*) Wilcoxon Matched-Pairs Signed Rank Test between all possible pairs of variables.

(3) Kendall's Coefficient of Concordance between persons.

The following subroutines from the IBM System/360 Scientific Subroutine Package are incorporated: SRANK, KRANK, CHISQ, QTEST, UTEST, TWOAV (modified), WTEST, RANK, TIE, ARRAY.

Language and Configuration The program is coded in FORTRAN IV for use on an IBM 7040 computer with 32K words of core memory. One scratch tape is required. A maximum of 200 subjects and 45 variables can be handled by the program in its present form.

Timing This will vary with the machine and the amount of data. Fifteen jobs of three variables and 35 to 40 subjects were run in less than 2½ min. on a machine with off-line printing.

Input, Options, Output The input can be from cards or from magnetic tape and data can be either ranked or unranked (a ranking subroutine is included). Transformations can be made on variables, and new variables can be generated as functions of input variables by adding appropriate FORTRAN statements into a data-transformation subroutine.

The data are preceded by a title card, a parameter card specifying the amount of data and the types of statistics desired, and a variable format card. Certain of the significance tests require an additional parameter card to be read after the data.

Any number of jobs may be stacked on a given run and execution is terminated by a blank card following any job. The output varies with the statistics desired.

Availability Listings and copies of the source deck are available from the Division of Educational Research Services, University of Alberta, Edmonton, Canada.

References

SIEGEL S. (1956). *Nonparametric Statistics for the Behavioral Sciences.* New York: McGraw-Hill.

IBM TECHNICAL PUBLICATION DEPARTMENT. *Application Directory, System/360 Scientific Subroutine Package.* New York: White Plains.

Factor Scores Programs. By H. J. HALLWORTH and ANN BREBNER, School of Education, The University of Birmingham, and College of Education, The University of Calgary, Alberta.

Description A program using the regression method to estimate factor scores forms part of a system of programs already abstracted in the Computer Section (Hallworth and Brebner, 1966). The present program is complementary in that it employs certain formulas for component scores derived by Kaiser (1962). It may therefore be used to obtain scores from principal components, or from components which have been rotated to either an orthogonal or an oblique position.

The maximum number of variables and components for current versions of the program is 50, but the number of subjects whose data can be analysed is virtually unlimited.

Computers, Languages The program is available in Autocode and FORTRAN versions suitable for running on the KDF–9 or Atlas computers. For the Autocode programs, input of program and data may be by means of either 5-hole paper tape with standard Autocode characters, or by cards. For the FORTRAN programs, input of both program and data is by cards. All current versions use a variable format facility.

FORTRAN versions of the program will also be made available for the IBM 360 series of computers.

Data Format All versions of the programs are self-regenerating, the number of sets of data being indicated at the beginning of a run on the computer.

The words COMPONENTS, or ORTHOGONAL FACTORS, or OBLIQUE FACTORS are used to indicate the type of scores required. Further control words regulate the type of output: scores may be punched on tape or cards, or printed.

It is always necessary to input a matrix of principal components, the means and standard deviations of all variables for which scores are required, and the subjects' scores on these variables. In the case of rotated factors, a matrix of loadings on these is also required. For OBLIQUE FACTORS it is also necessary to use the matrix of correlations between the factors.

The program does not require the use of all the principal components of a correlation matrix.

Computing Times Times will vary considerably according to the language and computer used. In most cases the actual computation time will be relatively insignificant and the over-all time will be dependent primarily on the amount of input and output required.

Availability Copies of the program tapes, card decks, listings and a full description of the program are available from the authors.

References

HALLWORTH, H. J. and BREBNER, ANN (1966). Systems of computer programs for multivariate anaysis. *Br. J. math. statist. Psychol.* **19**, 137–138.

KAISER, H. F. (1962). Formulas for component scores. *Psychometrika* **27**, 83–87.

PUBLICATIONS OF THE BRITISH PSYCHOLOGICAL SOCIETY

The British Journal of Psychology (about 512 crown quarto pages per volume), *The British Journal of Medical Psychology* (about 350 crown quarto pages per volume), and the *British Journal of Social and Clinical Psychology* (about 320 crown quarto pages per volume) appear in quarterly parts, the subscription price of each, payable in advance, being £4 10*s*. per volume post-free ($16.50 in the U.S.A.). Subscriptions to any of these journals may be sent to any bookseller, or subscription agent, or to the Cambridge University Press, Bentley House, 200 Euston Road, London, N.W.1. American enquiries should be addressed to the Cambridge University Press, American Branch, 32 East 57th Street, New York, N.Y. 10022.

The British Journal of Mathematical and Statistical Psychology is issued twice a year, in May and November, in volumes of about 288 pages. The subscription rate, payable in advance , is £3 10*s*. per volume post free ($13 in the U.S.A.). Subscriptions should be sent to the Society's Agents, Taylor and Francis, Ltd., Red Lion Court, Fleet Street, E.C.4; or to any bookseller.

The British Journal of Educational Psychology is issued three times a year, in February, June and November. The subscription is £2 a year post free, payable in advance, or £1 each number. The subscription price in the U.S.A. is $4.20. Subscriptions should be sent to the Manager, B.J.E.P. Department, Methuen and Co. Ltd., 36 Essex Street, Strand, London, W.C.2.

A series of *Monograph Supplements* is issued in connexion with *The British Journal of Psychology* but is not included in the subscription for that journal. Orders may be sent to any bookseller, or the Cambridge University Press, 200 Euston Road, London, N.W.1, from whom details of these publications may be obtained.

MANUSCRIPTS

for *The British Journal of Psychology* should be sent to
Professor A. Summerfield, Department of Psychology, Birkbeck College, Malet Street, London, W.C.1;

for *The British Journal of Medical Psychology* should be sent to
Dr. T. Freeman, Lansdowne Clinic, 4 Royal Crescent, Kelvingrove, Glasgow, C.3;

for *The British Journal of Social and Clinical Psychology* should be sent to
Mr. Michael Argyle (Social Psychology Editor), Institute of Experimental Psychology, 1 South Parks Road, Oxford, *or* to
Dr. R. R. Hetherington (Clinical Psychology Editor), 6 Abercromby Square, Liverpool 7;

for *The British Journal of Mathematical and Statistical Psychology* should be sent to
Professor R. J. Audley, Department of Psychology, University College London, Gower Street, W.C.1;

or *The British Journal of Educational Psychology* should be sent to
Mr. L. B. Birch, The Institute of Education, The University, Sheffield 10.

for *The Monograph Supplements* should be sent to
Professor I. M. L. Hunter, Department of Psychology, The University, Keele, Staffs., whose decision regarding the publication of all Monograph Supplements will be final. No MS. can be considered for publication in any year unless it is in the hands of the Editor by 1 September of the preceding year.

SEPARATES

Contributors receive on application fifty copies of their papers free. Additional copies of separates may be obtained in units of fifty copies.

SUBSCRIPTIONS TO THE BRITISH PSYCHOLOGICAL SOCIETY

The subscription for Fellows, Associates and Graduate Members is £8 10*s*. (£7 10*s*. if paid by Banker's Order). They are entitled to purchase any or all of the following Journals at the special rates of £1 10*s*. per Journal per annum: *The British Journal of Psychology*, *The British Journal of Medical Psychology*, *The British Journal of Social and Clinical Psychology*, *The British Journal of Mathematical and Statistical Psychology* and *The British Journal of Educational Psychology*.

Information on membership and activities may be obtained from the Secretary, The British Psychological Society, Tavistock House South, Tavistock Square, London, W.C.1.

0 DEC 1968

THE BRITISH JOURNAL OF MATHEMATICAL AND STATISTICAL PSYCHOLOGY

Volume 20, *Part* 1, 1967

CONTENTS

THOMAS, E. A. C.
Reaction-Time Studies: The Anticipation and Interaction of Responses *page* 1

BAINBRIDGE, LISANNE
A Discrete Predicting Controller 31

RODGER, R. S.
Type I Errors and their Decision Basis 51

LOCKHART, R. S.
The Assumption of Multivariate Normality 63

STRINGER, PETER
Cluster Analysis of Non-Verbal Judgements of Facial Expressions 71

PEEL, E. A.
A Method for Investigating Children's Understanding of Certain Logical Connectives used in Binary Propositional Thinking 81

CONSTANTINESCU, PAUL
A Method of Cluster Analysis 93

NOTES AND CORRESPONDENCE
Comparison of Factors by Ahmavaara's Method 107

BOOK REVIEWS 111

COMPUTER PROGRAM ABSTRACTS 125

Printed in Great Britain by
Taylor & Francis Ltd., Red Lion Court, Fleet Street, London E.C.4

ocie
ssue
Library R 8/

Volume 20, Part 2 *November 1967*

THE BRITISH JOURNAL OF MATHEMATICAL & STATISTICAL PSYCHOLOGY

EDITED BY R. J. AUDLEY

WITH THE ASSISTANCE OF CYRIL BURT

TAYLOR & FRANCIS LTD
RED LION COURT, FLEET STREET, LONDON, E.C.4

40s. net $7.45

Book Review Editor: A. R. JONCKHEERE

Abstracts Editor (Computer Programs): H. J. HALLWORTH

Editorial Board:

C. Banks
M. S. Bartlett
G. H. Bower
D. E. Broadbent
J. Brown
V. R. Cane
L. J. Cronbach
C. I. Howarth
A. R. Jonckheere
D. N. Lawley
R. D. Luce
S. Messick
E. A. Peel
L. S. Penrose
J. Fraser Roberts
W. Stephenson
A. Stuart
P. E. Vernon

88
7.70
Psy. R

PUBLICATIONS OF THE BRITISH PSYCHOLOGICAL SOCIETY

The British Journal of Mathematical and Statistical Psychology is issued by the British Psychological Society. Details of subscriptions to this and other Journals of the Society are given inside the back cover.

Papers for publication and books for review should be sent to the Editor, Professor R. J. Audley, Department of Psychology, University College London, Gower Street, W.C.1.

Abstracts of computer programs should be sent to Mr. H. J. Hallworth, Department of Education, University of Birmingham, Birmingham 15. Articles on computers, computer languages and programming will also occasionally be published and should be addressed to Mr. Hallworth.

PREPARATION OF PAPERS

Contributors are asked to send their papers in a form which is ready for submission to the printer: that is to say, the arrangement of the material should follow that observed in a current issue of this *Journal*. Each article should be headed with a brief abstract. Each section should have a short cross-heading, in capitals; and the whole should conclude with a summary embodying the main conclusions reached. Special attention should be paid to such details as correct spelling, grammar, capitalization, numbering, etc. In preparing manuscripts and in correcting proofs authors should conform, so far as possible, with the 'Recommendations to Authors' set out in *The Printing of Mathematics* by T. W. Chaundy, P. R. Barrett, and Charles Batey (Oxford University Press, 1954, ch. 2, pp. 21–73), or to those given in the paper on 'Setting Mathematics' (Arthur Phillips, *Monotype Recorder*, 60, iv, 1956).

Owing to the present conditions for printing, the cost of setting up tables and algebraic formulae which require hand composition is very much greater than the cost of machine composition. Contributors are therefore advised to arrange their material so that it can be printed as economically as possible. Often numerical results and algebraic formulae can, with a little simplification, be readily adapted for purposes of machine composition. Authors must ensure that formulae are clear, particularly when these contain subscripts.

Tables, photographs and figures which are costly to print should be used only when they save considerable space or are essential for effective presentation; and in no case should both a figure and table be used to convey the same information. Each table or figure should be numbered and on a separate sheet with its approximate position indicated in the typescript.

Drawings should be in Indian ink on heavy unruled paper. Lettering on graphs, etc. is best indicated in a tracing placed over the diagram. Diagrams are normally reduced in reproduction and allowance must be made for a corresponding reduction in the thickness of lines and the spaces between them. The numbered legend of a figure should be typewritten on a separate sheet of paper, not lettered on the drawing.

The object of the caption for a table or the legend for a figure should be to make the table or the figure comprehensible without reference to the text.

References in the text should be indicated by dates in parentheses after the authors' names and correspond to an alphabetical list at the end of the paper. The list must follow the layout illustrated in the three following examples:

ALLPORT, G. W. & VERNON, P. E. (1931). A test for personal values. *J. abnorm. soc. Psychol.* **26**, 231–248.

GALTON, F. (1883). *Inquiries into Human Faculty and its Development*. London: Macmillan.

RÉVÉSZ, G. (1951). Le problème du génie. *Année psychol.* **50**, 83–96.

World list abbreviations are used for journals. The Oxford Dictionary is taken as the authority on spelling. For other general guidance, authors are referred to *The Journal of Physiology: Suggestions to Authors*. Cambridge University Press. Price 2*s*.

In future authors will receive one complete slip-proof of their papers, but no page proof. They will be expected to pay the cost of all corrections for which they themselves are responsible.

All manuscripts should be submitted in duplicate.

PROGRAM ABSTRACTS

An abstract of a computer program, normally of 300 to 500 words in length, should state the sequence of standard mathematical or statistical operations that are performed, the language in which the program is written, the approximate running time and should either give an address from which the program can be obtained (in tape or card form) or state where data may be sent for processing. Authors are referred to a current issue of the *Journal* for examples of an acceptable abstract. In most cases, the abstract should be accompanied by an actual program in the form of a deck of cards or a paper tape together with the appropriate listing information, also any other instructions which are necessary for the preparation of data or the running of the program.

Diary No........................
Date..............................
File No...........................
Bureau Edni. Psy. Research

THE BRITISH JOURNAL OF MATHEMATICAL & STATISTICAL PSYCHOLOGY

EDITED BY R. J. AUDLEY

WITH THE ASSISTANCE OF CYRIL BURT

VOLUME XX

TAYLOR & FRANCIS LTD
RED LION COURT, FLEET STREET, LONDON, E.C.4

Suggestions to Authors

A new publication of The British Psychological Society

This booklet contains advice on the preparation of papers for submission to journals published by the Society and gives general information as well as notes on the arrangement of articles; preparation of manuscripts; mathematical and statistical formulae; recommended abbreviations and symbols; recommended spelling of commonly used terms; proof correction; a list of periodicals and their abbreviations and notes on copyright arrangements.

Available from

THE CAMBRIDGE UNIVERSITY PRESS
BENTLEY HOUSE, P.O. Box 92, 200 EUSTON ROAD, LONDON, N.W.1

(5/- per copy, post free)

A SYSTEM OF COMPUTER PROGRAMS

FOR USE IN PSYCHOLOGY AND EDUCATION

'A SYSTEM OF COMPUTER PROGRAMS' (by Hallworth and Brebner, price 10/-) is usable even by people who have no experience with computers. For the purpose of analysing data using means and standard deviations and correlations, or a principal components analysis, with Varimax or Promax rotations (and second order factors) and factor scores; or Z-scores, partial correlations, multiple correlations or regressions, this system provides a set of programs which can be used on a computer which uses ICI C-Autocode, ICI K-Autocode, Mercury Autocode or Fortran.

The handbook tells you exactly how to prepare your data for the computer, and how to interpret the figures that come from the computer.

The HANDBOOK is available from the British Psychological Society, Tavistock House South, Tavistock Square, London, W.C.1 (please enclose 10/- with order.)

PROGRAM TAPES, CARD DECKS AND LISTINGS are available at cost from System of Computer Programs, Education Department, University of Birmingham.

Vol. 20
Part 2
129–162

The British Journal of Mathematical and Statistical Psychology

November
1967

TEMPORAL SUMMATION AND ABSOLUTE BRIGHTNESS THRESHOLDS

By T. Shallice

University College London

Three normative models based on the Change Detector analogies of Shallice (1964) are investigated to see how well they account for the integration of input over time that occurs in absolute brightness threshold experiments. Of the three, the geometric moving average receives no support. The perceptual moment and moving average models give about equally satisfactory but not perfect fits for short flash results, but the perceptual moment model is much more satisfactory for long flash results.

The analysis takes into account variations in information transmission times, noise in the visual system and the quantal nature of the light input. The locus of temporal integration is discussed and a central locus preferred, following Treisman (1966).

1. Introduction

The two major modern theories of detection—signal detection theory and the fixed threshold theory—both assume that some internal representation of the stimulus is compared with a criterion (see Luce, 1963). If and only if the representation is greater than the criterion is the stimulus detected. The two theories do not discuss how the central representation is produced. However, where the theories have been applied to particular senses this has been done, e.g., in vision by Barlow (1956, 1957, 1958), Nachmias and Steinman (1963), Treisman (1964 b, 1966) and Baumgardt and Smith (1965). Yet one important question which even these theorists either ignore or answer by arbitrary assumptions is over what period of time information is integrated to give rise to the central representation upon which the decision will be based.

In a previous paper (Shallice, 1964) the present author argued that in a typical threshold experiment the subject tries to detect any change in the stimulus so that the situation is analogous to that of a quality control statistician who has to detect whether or not the output of a manufacturing process has changed. Three models of temporal integration in detection were produced by analogy with the statistician's methods and tested in three situations: low frequency absolute thresholds for hearing, absolute thresholds for vision, and the perception of causality. The third of these tests was later shown (Shallice, 1965) to be based on additional faulty assumptions, whilst the second test was based on an over-restrictive assumption. The purpose of the paper is to attempt to remedy this last fault by considering absolute brightness thresholds in more detail.

Absolute brightness thresholds are a suitable experimental situation for testing theories of temporal integration. Both the physical and the proximal stimulus can be described mathematically and this is the situation where two of the complicating factors normally involved in studying the visual system—adaptation and lateral inhibition (see Barlow, Fitzhugh and Kuffler, 1957)—can be most safely ignored in the early stages of theorizing. In addition, it is better understood, both physiologically and psychologically, than most perceptual situations.

One objection to the use of absolute brightness thresholds is that the traditional explanation of Bloch's Law—that the threshold energy of a flash is independent of its duration for flashes lasting less than a critical duration—is that it results from integration of information in the retina (see, e.g., Graham and Margaria, 1935). Since Bloch's Law is the best known phenomenon of temporal integration in vision, this indicates that temporal integration in this situation is a peripheral effect and so may not be typical. This objection is not, however, very strong. In the first place, temporal integration in other detection situations might be peripheral as well. Secondly, suggestions have been made quite frequently in recent years that in this and similar visual situations the locus of temporal summation is central and not peripheral (e.g. Johnson, 1958; Boynton and Siegfried, 1962; Blackwell, 1963; Kahneman and Norman, 1964; and Treisman, 1964 b, 1966). Finally the absolute threshold, being the most basic phenomenon in visual perception, is important in its own right. It might seem that ignorance of where temporal integration occurs is a severe handicap for producing models, but the models discussed in this paper are normative, in the way that signal detection theory is, and no anatomical locations are specified. Whilst they were originally produced with a central interpretation of them in mind, two can be interpreted in peripheral terms. The whole question of the anatomical locus of temporal integration will be discussed in Section 10.

It will be assumed in common with most other theories in this field that information is integrated over time in a unit which will be called an accumulator. An analogous approach has had success in explaining reaction time results (e.g. McGill, 1963; LaBerge, 1962; Laming, 1968; Shallice and Vickers, 1964). One theory in the area does not contain this assumption: that of Blackwell (1963) who assumes very reasonably that the central effect of the light absorbed in time δt does not just last for time δt, but due to varying transmission lags in the system, lasts for considerably longer. So light absorbed at different times produces central effects at the same later instant of time. This, too, is analogous to a useful theory in the area of reaction times—that of Raab (1962). Yet the absorption of light is a discontinuous process and information from eye to brain is transmitted in discrete pulses. Hence some integrating mechanism is necessary, if only one with a very short time constant such as a typical nerve cell excitatory post-synaptic potential (EPSP). However, due to the inherent variability of biological material, and particularly the known variability of the visual system's response (e.g. Fitzhugh, 1957), it will also be assumed that the transmission lags from absorption to the accumulator are variable; so removing the over-restrictive assumption of Shallice (1964). The distribution of transmission lags will be called the lag distribution. For most of the paper the lag distribution will be considered as a confounding variable obscuring the effect of the integrating mechanism. In the Discussion, Blackwell's theory, where the two roles are reversed, will be considered further.

2. CHANGE DETECTOR MODELS

Before discussing the models, the assumptions common to all three will be specified. Figure 1 gives a flow diagram of these common properties. Some of these assumptions are made to fit widely supported physiological theories; others, on issues where there seems little or no physiological evidence, so as to make the models as mathematically tractable as possible. As in all model building it is necessary to compromise between usefulness and correctness: too many assumptions make prediction very difficult and corroboration relatively meaningless, too few make corroboration impossible. Here, most of the assumptions err on the side of simplicity: greater specificity would require

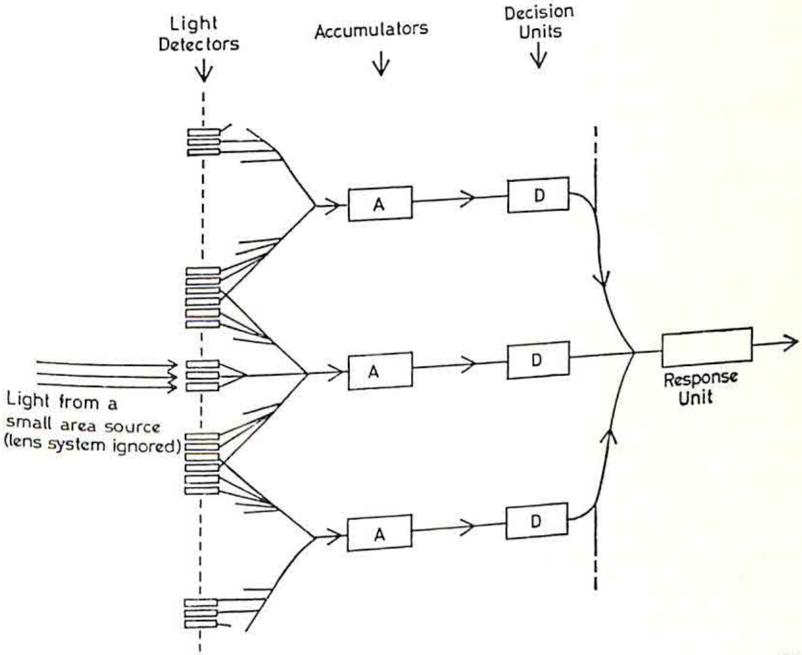


FIGURE 1. The flow diagram of a section of the ideal flash detector system used in this paper. The light detectors correspond to rods. The decision units are in one of two states 'Yes' or 'No' and can operate according to the principles of signal detection theory or of the fixed threshold theory. The response 'No' occurs only if all the decision units output 'No', otherwise 'Yes' occurs. Normally in this paper very small diameter flashes are considered (e.g. 2′ or 3′) so that only one Accumulator/Decision system is relevant.

greater physiological knowledge. Some of the predictions made will not depend on all of these assumptions and the relation between the assumptions and the predictions will be clarified in the Discussion.

1. *In the transmission of information from the stimulus to its internal representation, temporal integration occurs at one level only.*

This precludes, for instance, a system in which a retinal ganglion cell integrates input and its output is integrated by a higher mechanism.

2. *Each photon arriving at the eye has an equal chance, v, of being absorbed and so triggering an input to the accumulator of a constant number of impulses in a short burst, after a lag which is independent of the lags of other inputs.*

For simplicity it will be assumed that this constant number is unity, since this makes the lag distribution easier to conceptualize; the deductions to be made are, however, compatible with any constant. Treisman's (1964 b, 1966) theory about the variability in burst number will be discussed in Section 10.

The idea that the absorption of a single photon in a rod can lead to neural activity is a necessary consequence of the statistical arguments of Hecht, Shlaer and Pirenne (1942) (see Brindley, 1960 b).

3. *An accumulator receives input only from the field of one rod ganglion cell. Decisions are based on the output of individual accumulators.*

This is to account for Ricco's Law (see Brindley, 1960 b), and for the fact that Bloch's Law breaks down more sharply for larger fields (Bouman and van der Velden, 1947; Barlow, 1958; Baumgardt and Hillman, 1961). This can be explained by probability summation over different accumulators (see Section 9).

4. *In addition to the photon-produced impulses arriving at an accumulator there are assumed to be ' noise ' impulses—probably produced by the spontaneous degeneration of rhodopsin. These are indistinguishable from the photon-produced impulses and occur at a rate corresponding to x absorbed photons per accumulator.*

The assumption of noise is supported by microelectrode recordings (Kuffler, Fitzhugh and Barlow, 1957) and by psychophysical results (Barlow, 1956, 1957; Nachmias and Steinman, 1963). Barlow's (1957) conclusions have been challenged by Baumgardt and Smith (1965, 1966), but Nachmias (1966) has produced a plausible counter-explanation. If Baumgardt and Smith are correct, x, which can be estimated from certain figures in this paper, should be zero. So to include the assumption increases the generality of the predictions.

Three models for the operation of the accumulator will be compared; all are based on statistical quality control analogies and seem the simplest conceivable. They were discussed in Shallice (1964) and will therefore be only briefly described here:

(i) *Geometric Moving Average Model* (GMA)

In this model the accumulator suffers a constant leak such that any impulse in the accumulator has a constant choice of decaying in a given time interval. Its behaviour can therefore be described by an immigration–death process.

Zwislocki (1960) has suggested a model for auditory temporal summation which is a continuous analogue of this model. Moreover, Ten Hoopen (1966) has used the identical mathematical model to explain the behaviour of EPSP in nerve cells, so that one obvious physiological representation of the model is summation of EPSP's in, say, a retinal ganglion cell. Another would be some analogy with the decay theory of memory.

(ii) *Moving Average Model*

On this model, the input is integrated totally over a certain period (P) and then discarded. Formally, the content of the accumulator at time t, y_t, is:

$$y_t = \lim_{\Delta t \to 0} \frac{1}{\Delta t} \int_{t-P}^{t} z_t \, dt,$$

where $z_t = 1$ for intervals of length Δt, each of which contains the arrival time of one impulse,
$= 0$ otherwise.

Its major disadvantage is that a complete record of the time the input arrived must be kept for the time P. The model is therefore much less economic in memory space than any of the others; in compensation it is faster. In certain cases it could have a retinal representation and it is in fact mathematically very similar to the traditional explanation of Bloch's Law of summation over a fixed interval (see Bartlett, 1965).

(iii) *Perceptual Moment Model*

On the perceptual moment model the time axis is divided into discrete 'moments' of length D and y_t is only calculated at the end of moments. If y_m is the output of moment m, then

$$y_m = \lim_{\Delta t \to 0} \frac{1}{\Delta t} \int_{D(m-1)}^{Dm} z_t \, dt.$$

This model is formally similar to but functionally different from the model for general brain organization proposed by Stroud (1955) but criticized by Shallice (1964, 1965), Allport (1966) and Harter (1967).

A modification of this model is to allow y_t to be calculated at all points in time and not just at the end of moments. This, however, gives exactly the same predictions as the perceptual moment model in absolute threshold situations. It is similar to the model for reaction times suggested by Audley (1963).

(iv) *Sequential Sampling Model*

All these schemes make use of mainly recent input and so lose information. An alternative model which takes account of all information after some arbitrary starting point $t = 0$ is obtained by letting

$$y_t = \lim_{\Delta t \to 0} \int_{0}^{t} (z_t/\Delta t - x) \, dt.$$

This model gives the prediction that if one does not know when the flash will occur the chance of detection will depend, to a close approximation, only on the energy in the flash and not on its duration. This is not the case, so the model will not be considered further.

3. The Conditions for Testing the Models

Experimental work on temporal summation and absolute brightness thresholds may be divided into two parts according to the length of time the flash or flashes last. First, there is the work done on short flashes mainly to test Bloch's law of complete temporal summation. Experiments since 1940 have included Bouman and van der Velden (1947), Long (1951), Bouman and van den Brink (1952), Davy (1952), Rouse (1952), Barlow (1958), Baumgardt and Hillman (1961), Blackwell (1963) and Sperling and Jolliffe (1965). Second, there is the work on the detection of long flashes, say longer than one second. Although a number of experimenters have worked on flashes lasting up to about one second, the only work on long flashes known to me, which is sufficiently detailed to be of use for testing quantitative predictions is that of Bouman and van der Velden (1947).

The reason for making this division into short and long flashes is connected with the difficulty of dealing with the lag distribution. If the flash is long compared with the spread of the lag distribution, then since each transmission lag was assumed independent of all other transmission lags, a Poisson stimulus input becomes a Poisson input to the accumulator except at the beginning and end of the flash duration. This follows from a general theorem on the random displacement of events in point processes (see Cox and Miller, 1965, p. 365). However, for a short flash, even if the stimulus input is Poisson (as we know it is physically), the stochastic nature of the input into the accumulator will depend upon the nature of the lag distribution. For example, if the lag distribution were rectangular with width equal to the duration of the stimulus the density of the arrival of impulses at the accumulator would be triangular, not constant as a Poisson input should be. So in order to obtain predictions for short flash experiments it is necessary to make a further assumption about the nature of the lag distribution. It is not necessary to make this assumption when dealing with the long flash predictions, which makes the long flash results better for discriminating between models.

These theories can at present only be tested psychologically. However, before accepting the identification of the output of the accumulator with some measurement of perception, it is necessary to consider whether other factors could affect the experimental results. Possibilities are the psychophysical method used, the expectation of the subject and eye movements.

Of the experimenters mentioned earlier in this section, Bouman and van der Velden, Bouman and van den Brink and Baumgardt and Hillman all used the constant method, and Blackwell used the forced choice method. The others used either the method of limits or the adjustment method, which are less satisfactory (see Osgood, 1953), but it is not obvious how systematic alterations to the threshold could occur.

If, as Barlow (1956, 1957) suggested, absolute brightness thresholds should be considered as signal/noise discriminations, then the subject could lower his criterion at the appropriate time, if he knew exactly when the stimulus was due. This has been shown to occur with absolute thresholds for electric phosphenes and with difference thresholds for both light and sound intensity by Howarth and Treisman (1958) and Treisman (1964 a). If this effect occurred for absolute brightness thresholds and if the subject knew exactly when the stimulus was due to start, a subject could lower his criterion for a short flash and so distort the comparison of short with long flashes. Bouman and van der Velden give no details of their experimental procedure, so it is not possible to assess whether this artefact could have affected their results.

A third possibility suggested to me by A. H. Gregory is that if the subject does not fixate the fixation point rigidly, all the light from a flash may not fall on the same part of the retina. Then the threshold may be raised as the light may not fall in the field of the same ganglion cell. It is known that almost complete areal summation occurs up to at least 10′ in the periphery (Bouman and van der Velden, 1947; Barlow, 1958; Baumgardt and Hillman, 1961), so that only eye movements of that order will affect the results. However, Riggs, Armington and Ratliff (1954) showed that when subjects are fixating, eye movements are much smaller than this. In a 500 msec period, the image has only a

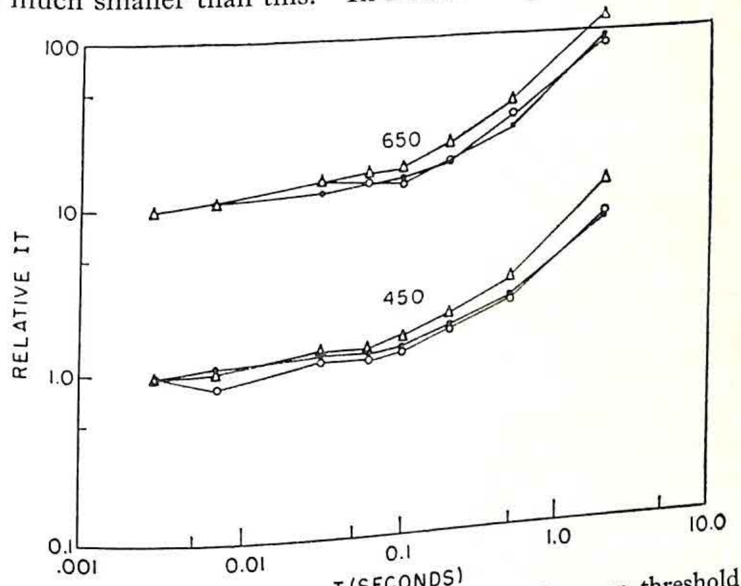


FIGURE 2. Sperling and Jolliffe's results for the relation between threshold energy and flash duration using peripheral stimuli of three diameters (3° (triangles), 45′ (open circles) and 4·5′ (filled circles)) and of two wavelengths (450 mμ and 650 mμ). The curves have been moved vertically so as to be equal at the shortest duration. (From *J. opt. Soc. Amer.*, 1965, **55**.)

50 per cent chance of moving more than 2′. Since summation times are of the order of 100 msec, fixation difficulties should have a negligible effect on the results. However, for foveal thresholds where areal summation is over a much smaller area (see Brindley (1960 b)) this artefact could be important. This is one reason why this paper deals only with peripheral thresholds and, except in Section 9, with flashes of very small area.

Experiments on the effect of duration on the absolute threshold normally show their results in the form of a relation between the threshold energy ($N=IT$) and the flash duration (T). This curve, which will be called the N/T curve, typically shows an upward kink in the region of 100 msec (see Figures 2 and 3); d^2N/dT^2 is high in this region. This kink is equivalent to the breakdown in Bloch's law with increasing duration. Since the interpretation of short flash results is complicated by the unknown nature of the lag distribution, the question to be investigated for these results is whether the models can provide a reasonable explanation of the kink. There appears to be no problem for the perceptual moment model and the moving average model, provided one assumes an integrating period of about 100 msec, since the effect of the flash cannot be confined to one integrating period for longer flashes. This is an appropriate assumption as it accords with results from low frequency thresholds (see Shallice (1964, 1965)). In fact, for the perceptual moment model the interpretation of short flash results is more complex but still satisfactory as will be seen in a later

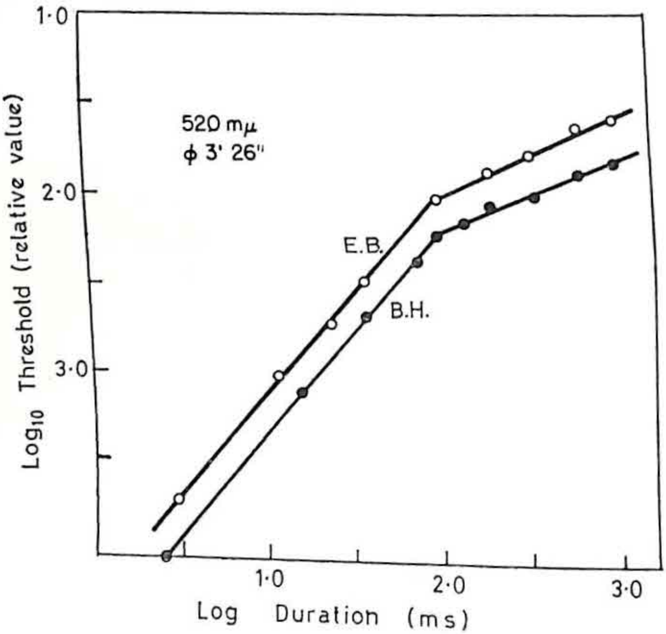


FIGURE 3. Baumgardt and Hillman's results for the relation between flash intensity and duration for small diameter flashes (3′ 26″). Note that the intensity is plotted increasing downwards. (From *J. opt. Soc. Amer.* 1961, **51.**)

section. For the GMA model no reasonable explanation of the kink has so far been produced, as will be shown in the next section.

Since the complication introduced by the lag distribution does not affect long flash predictions, for this condition the models will be examined to see how closely they predict existing frequency of seeing curves, which is a much more stringent test than that of explaining the short flash kink.

4. SHORT FLASH THRESHOLDS—GMA MODEL

It has previously been shown, using a continuous analogue of the GMA (Shallice, 1964), that it is not possible to explain the short flash kink if it is assumed that the lag distribution has zero variance.

This may be because the GMA is not a suitable model for visual thresholds or just because an unreasonable value was chosen for the lag variance. In this section, how the GMA behaves when a more realistic value of the lag variance is used will be examined in order to assess whether the previous result was just a consequence of the unreasonable lag variance used. This entails making some estimate of the lag distribution. To do this the theory which has been of most help is that of Mueller (1954). He proposed that every absorbed photon immediately changes one molecule of the absorbing substance to a second substance. This second substance subsequently goes to a third state and in so doing triggers an impulse in the bipolar cell and hence the ganglion cell, provided it is not in a refractory state. Mueller assumes that this second change of state is analogous to radioactive decay, in that the time between the first and second changes of state can be described by an exponential distribution. He shows that the theory can account for previous ganglion cell recordings of the Hartline (1934) type without assuming, as Hartline did, temporal integration in the *Limulus* ommatidium.

Mueller's theory satisfies the present axioms and it suggests that if most of the variance is due to delays early in the system, then the lags can be described by an exponential distribution plus a constant (τ). Using this as the lag distribution, it is possible to obtain predictions for the N/T curve on the GMA, provided the continuous analogue of the GMA is used. It has not yet proved possible to derive the N/T curve for the stochastic GMA.

If $L^*(t)\Delta t$ is the chance of any receptor cell producing an impulse in the time $[t, t+\Delta t]$ on Mueller's theory, and $L(t)\Delta t$ the chance of an impulse arriving at the Detector on the present model during this interval, then if x is the rate of production of noise impulses:

$$L(t) = L^*(t-\tau) + x. \tag{1}$$

The last term arises from Axiom 4. Further, from the postulates of the GMA, using a continuous analogue, we have that, if $y(t)$ is the number of impulses in the accumulator at time t,

$$\frac{dy}{dt} = L(t) - by. \tag{2}$$

Now Gregory (1963) has shown that for a semi-infinite flash starting at $t=-\tau$:

$$L^*(t)=q_1-(q_1-q_0)\,\mathrm{e}^{-c(t+\tau)}, \tag{3}$$

where q_0, q_1 are the average number of photons being absorbed before and after the flash started and c is Mueller's constant. In our example $q_0=0$, $q_1=\nu I$, therefore

$$\frac{dy}{dt}=x+\nu I(1-\mathrm{e}^{-ct})-by$$

so that

$$y=(x+\nu I)/b-\nu I\,\mathrm{e}^{-ct}/(b-c)+K\,\mathrm{e}^{-bt}, \tag{4}$$

where K is a constant.

The steady state level of the accumulator with no flash acting is given when $t\to\infty$ and $I=0$,

$$\text{i.e. } y=x/b.$$

Therefore, when the flash is about to start, $y=x/b$. Hence from eqn. (4),

$$y=\frac{x}{b}+\nu I\left(\frac{1}{b}-\frac{1}{b-c}+K\right)$$

when $t=0$. So that generally,

$$y=(x+\nu I)/b-\nu I(b\,\mathrm{e}^{-ct}-c\,\mathrm{e}^{-bt})/b(b-c). \tag{5}$$

Now if the flash lasts for a time T, then until $t=T-\tau$, eqn. (3) describes the generation of impulses at the receptors. Also, since the second term decreases with increasing t, the maximum value of $y(t)$ in the range $0\leqslant t\leqslant T$ is attained at $t=T$. Therefore

$$y(T)=(x+\nu I)/b-\nu I(b\,\mathrm{e}^{-cT}-c\,\mathrm{e}^{-bT})/b(b-c). \tag{6}$$

When $t>(T-\tau)$, we are interested in the output of the receptor after the flash has ended. Gregory gives the following equation for this condition:

$$L^*(t)=q_0+(q_1-q_0)(1-\mathrm{e}^{-cT})\,\mathrm{e}^{-c(T+t+\tau)}.$$

From eqns. (1) and (2) we have

$$\frac{dy}{dt}=x+\nu I(1-\mathrm{e}^{-cT})\,\mathrm{e}^{-ct}-by.$$

Therefore

$$y=-x/b+\nu I(1-\mathrm{e}^{-cT})\,\mathrm{e}^{-ct}/(b-c)+K'\,\mathrm{e}^{-bt},$$

where K' can be obtained from eqn. (6).

Hence $y(t)=x/b+\nu I[(1-\mathrm{e}^{-ct})\,\mathrm{e}^{-ct}/c-(1-\mathrm{e}^{-bT})\,\mathrm{e}^{-bt}/b]c/(b-c)$.

Now for detection, $[y(t)]_{\max}\geqslant C$, a constant.

By differentiation it can be shown that $[y(t)]_{\max}$ occurs when

$$(1-\mathrm{e}^{-cT})\,\mathrm{e}^{-ct}=(1-\mathrm{e}^{-bT})\,\mathrm{e}^{-bt}.$$

Therefore

$$[y(t)]_{\max}=x/b+\nu I(1-\mathrm{e}^{-cT})^{b/(b-c)}/b(1-\mathrm{e}^{-bT})^{c/(b-c)}.$$

If I and T are varied so that $[y(t)]_{\max}$ is constant, the relation between T and $N(=IT)$ should be

$$N = EbT(1-e^{-bT})^{c/(b-c)}/(1-e^{-cT})^{b/(b-c)},$$

where $\nu E = C - x/b$ is constant and N is the flash energy.

In Appendix 1 it will be shown that whatever the values of E, b and c this relation is not a good fit to empirical N/T curves.

This therefore suggests that the previous finding that the GMA model could not explain the N/T curve was not just a result of choosing an inappropriate lag distribution. However, since the present analysis is not based on the proper stochastic GMA model but on a continuous analogue of it, the evidence is not entirely conclusive.

5. SHORT FLASHES—PERCEPTUAL MOMENT MODEL PREDICTIONS

In Section 3 it was mentioned that while it is obvious that the perceptual moment model will predict a kink in the N/T curve about where $T=D$, there are features of the N/T curve which are hard for the model to predict. U. Neisser (personal communication) has pointed out that Bloch's law itself cannot be predicted on the perceptual moment model. He says, " It seems to me that the familiar IT curves would be produced by a quantal (moment) mechanism only if one could be sure that each flash began at the beginning of a moment. But this cannot be ensured, hence the moment (instant) of incidence must be linearly distributed across the quantal interval. It is easy to show that in this case there will be a parabolic relation between the duration of the stimulus and the average energy it delivers which does not level off until two quanta are reached. The observed linear relations are thus refutations of the moment theory."

This is a very strong argument and if Bloch's law holds exactly it seems to be a decisive refutation of the perceptual moment model. However, the experimental evidence on this point is not unequivocal. Some experiments find it holding exactly; others only approximately. Before reviewing the evidence, it seems advisable to obtain an estimate of the perceptual moment model predictions. Neisser's parabolic relation holds for average energy, but to extend it to actual thresholds requires two assumptions—that there is no lag variance and that for any length of flash the frequency-of-seeing curve is a step function. It has not proved possible to obtain a quantitative prediction when lag variance is allowed for, but it has proved possible to relax the second assumption.

If there is little or no lag variance then, considering the input to the accumulator from a very short flash, the impulses should arrive at the accumulator close together in time and hence most likely within one moment. The number of impulses arriving within any moment has to be greater than a certain criterion for detection, but the number of impulses arriving during this moment for a given intensity of flash depends upon the Poisson variability of the number of photons absorbed and the Poisson variability of the dark noise. Hence, as predicted for similar reasons, and tested by Hecht, Shlaer and Pirenne (1942), van der Velden (1944) and Barlow (1956), the frequency of seeing curves for very short flashes should be Poisson sums (see Figure 4). This use of short flash results receives further support from an argument in Section 9.

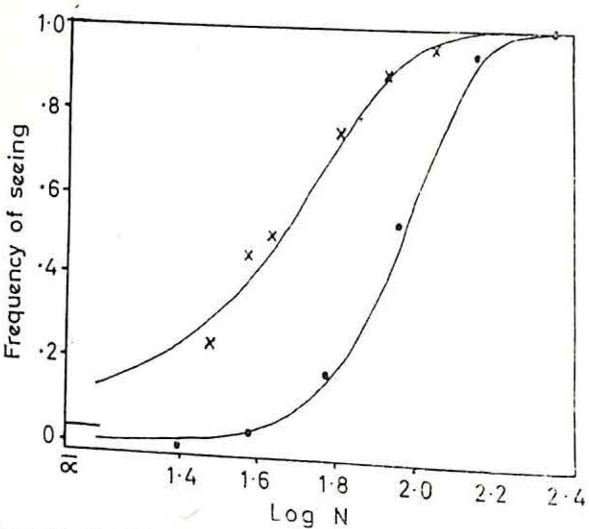


FIGURE 4. Data from Hecht, Shlaer and Pirenne (1942) (dots) and van der Velden (1944) (crosses) fitted by theoretical curves from Barlow's theory. (From *J. opt. Soc. Amer.* 1956, **46.**)

When one wishes to compare the energy of a longer flash necessary for a certain frequency of seeing with the energy of a short flash, it is necessary to consider what happens when the impulses from a flash do not all fall in one moment. Assuming no lag variance, then the chance that a flash of energy N all of whose impulses arrive in one moment will be seen is given by the frequency-of-seeing curves for a very short flash of energy N. Instead of the step function assumed by Neisser, it will be assumed that the frequency-of-seeing curves have the following form (where B, C are constants):

$$Q = \begin{cases} 0 & \text{if } \ln N \leqslant C, \\ \dfrac{\ln N - C}{B - C} & \text{if } C < \ln N < B, \\ 1 & \text{if } B \leqslant \ln N. \end{cases}$$

That this linear relation is a much better approximation than a step function can be seen from Figure 4.

Consider a flash of length T where $T \leqslant D$. Since it is assumed that there is no lag variance, the input to the accumulator lasts for time T. Now this may not be entirely within one moment, so assume that the larger of the two parts has length t. For an initial estimate of the relation between N and T it will be assumed that the flash is not detected in the smaller of the two parts. The probability of the flash being seen in the larger part is that of a very short flash of energy It being seen and is

$$\frac{\ln(It)-C}{B-C} \quad \text{if} \quad C<\ln(It)<B.$$

Now for a flash of length T, where $C<\ln(IT)<B$, the chance of its being seen, taking into account the variability in the instant of incidence is:

$$Q=\frac{D-T}{D}\,\frac{[\ln(IT)-C]}{B-C}+\frac{2}{D}\int_X^T\frac{\ln(It)-C}{B-C}\,dt, \tag{7}$$

where $X=\max[T/2, \exp(C)/I]$.

Now if $\ln(IT)>C+\ln 2$, then $X=T/2$ and from (7):

$$(B-C)Q-\frac{D-T}{D}[\ln(IT)-C]=\frac{2}{D}[t\ln(It)-t-Ct]^{T}_{T/2}.$$

Therefore $$(B-C)Q=\ln(IT)-C-T(1-\ln 2)/D. \tag{8}$$

Hence for constant frequency of seeing, the following relation must hold:

$$\log IT=K+0\cdot 13T/D, \tag{9}$$

where K is a constant.

This relation holds for $0<T<D$ provided:

(i) $\ln N<B$ when $T\to D$ and (ii) $\ln N-\ln 2>C$ when $T\to 0$.

From the results of the experiments mentioned above, $(B-C)$ seems to lie roughly between 0·6 and 0·8 $\log_{10}$ units.

Considering relation (ii),

$$\begin{aligned}\ln(IT)-C&>\tfrac{1}{2}(B-C)\quad \text{if } Q>\tfrac{1}{2}\\ &>\tfrac{1}{2}\ln 10\times 0\cdot 6\\ &=\ln 10\times\log 2\ (\text{using } \log 2=0\cdot 3010)\\ &=\ln 2.\end{aligned}$$

Hence relation (ii) holds provided $Q>0\cdot 5$.

From eqn. (8), $C+(B-C)Q=\ln(IT)-0\cdot 13\ln 10$. Therefore,

$$\begin{aligned}\ln(IT)-0\cdot 13\ln 10&=BQ+(B-0\cdot 6\ln 10)(1-Q)\\ &=B-0\cdot 6\ln(1-Q).\end{aligned}$$

So $$\ln(IT)<B \text{ if } 0\cdot 6\,(1-Q)>0\cdot 13\ln 10, \text{ i.e. } Q<0\cdot 78.$$

Hence both relations hold if Q lies between 0·5 and 0·78 approximately. Since the method of limits presumably gives the 50 per cent threshold and Bouman and van der Velden used the 60 per cent threshold, this range covers all the normally used criteria.

There are five modern experiments known to me which give N/T curves for short small area peripheral flashes. In one of these (Barlow, 1958) the results are presented graphically in a form in which it is not possible to tell whether the moving average model prediction of $N=\text{constant}$ for $0\leqslant T\leqslant D$ or eqn. (9) fits best. Another (Rouse, 1952) was principally concerned with for $0 \leqslant i \leqslant n$, of the rth moment of the distribution of the time to absorption from state i, then:

$$(\mathbf{I}-\mathbf{T})\mathbf{m}^{(r)} = r\mathbf{m}^{(r-1)} \ldots + (-1)^{(r+1)}\mathbf{m}^{(0)}.$$

Let μ be the rate of input to the accumulator $(=\nu I)$ and λ be the rate of decay, which is equivalent to the death rate. Then, if h is small,

$(\mathbf{I}-\mathbf{T})/h =$

$$\begin{bmatrix} \mu & -\mu & 0 & \cdot & \cdot & \cdot & \cdot & \cdot & 0 \\ -\lambda & (\lambda+\mu) & -\mu & 0 & \cdot & \cdot & \cdot & \cdot & 0 \\ 0 & -2\lambda & 2\lambda+\mu & -\mu & 0 & \cdot & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & \cdot & \cdot & \cdot & \cdot & 0 & -(n-2)\lambda & \mu+(n-2)\lambda & -\mu \\ 0 & \cdot & \cdot & \cdot & \cdot & \cdot & 0 & -(n-1)\lambda & \mu+(n-1)\lambda \end{bmatrix}$$

Consider $\mathbf{T}' = \begin{bmatrix} \mathbf{T} & \mathbf{C} \\ 0 & 1 \end{bmatrix}$ an $(n+1) \times (n+1)$ square matrix, where $\mathbf{C}$ is the $n \times 1$ column vector $[0, \ldots, 0, \mu]$. Now $\mathbf{T}'$ is a finite irreducible Markov chain with only one absorbing state and no other persistent states. Hence the probability of absorption $= 1$ (Feller, 1957).

Therefore, using Pike's formula for the mean absorption time,

$$(\mathbf{I}-\mathbf{T})\mathbf{m}^{(1)} = \mathbf{m}^{(0)} = 1.$$

If U_i is the mean number of units of time, h, from state i to absorption,

$$U_0 - U_1 = 1; \tag{10}$$

$$-i\lambda U_{(i-1)} + (i\lambda+\mu)U_i - \mu U_{(i+1)} = 1 \quad (1 \leqslant i \leqslant n-2); \tag{11}$$

$$-(n-1)\lambda U_{(n-2)} + (\mu + (n-1)\lambda)U_{n-1} = 1. \tag{12}$$

For the first $(n-1)$ equations, the difference eqn. (11) together with the boundary condition (10) has the solution

$$U_i = U_0 - \sum_{j=0}^{i-1} \frac{\lambda^j}{\mu^{(j+1)}h} \sum_{r=j}^{i-1} \frac{r!}{(r-j)!}.$$

Define U_n by the equation

$$-(n-1)\lambda U_{(n-2)}[(n-1)\lambda+\mu]U_{(n-1)} - \mu U_n = 1. \tag{13}$$

Then U_n satisfies the difference eqn. (11). Therefore

$$U_n = U_0 - \sum_{j=0}^{n-1} \frac{\lambda^j}{\mu^{(j+1)}h} \sum_{r=j}^{i-1} \frac{r!}{(r-j)!}.$$

But from the nth eqn. (12) as well as eqn. (13), we have $U_n = 0$. Hence

$$U_0(t) = \sum_{j=0}^{n-1} \frac{\lambda^j}{\mu^{(j+1)}} \sum_{r=j}^{i-1} \frac{r!}{(r-j)!} \quad \text{as } h \to 0. \tag{14}$$

Ten Hoopen (1966) has obtained by different means a result which reduces to this. He has also obtained an expression for the variance of the distribution, but since one term involves the solution of a difference equation it can only be solved for particular cases and cannot be used as a general relation. In the present case $\mu = \nu I + x$ where ν is the proportion of arriving photons absorbed by the rods, x is the rate of production of ' dark noise ' impulses and λ is the unknown constant of the GMA process.

There are two reasons why it is not possible to use this relation to test the GMA. First, consider how the distribution relating frequency of seeing to flash duration varies with changes in the intensity of the flash. Eqn. (14) shows how the mean of the distribution should vary, but all the experimental results show how a particular percentile of the distribution varies and one cannot be deduced from the other. An exception to this generalization is provided by the frequency-of-seeing curves found by Bouman and van der Velden (see Figure 6). However, their curves show how, for fixed duration flashes, flash energy and

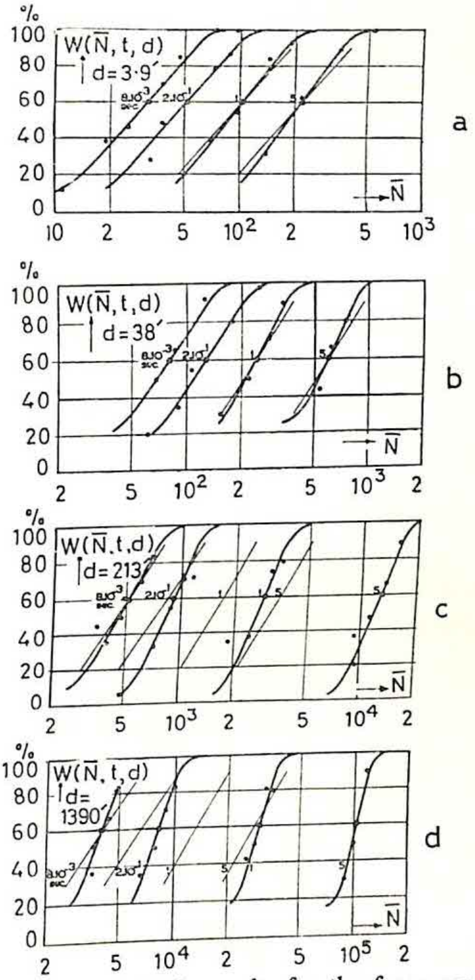


FIGURE 6. Bouman and van der Velden's results for the frequency-of-seeing curves for flashes of different diameters (d), durations (t) and energies (N). The thin lines are the predictions from their two quanta theory. (From *J. opt. Soc. Amer.* 1947, 37.)

hence flash intensity is related to frequency of seeing, but what is required is how, for fixed intensity, flash duration is related to frequency of seeing. This could be roughly estimated from their results, but the mean depends upon all parts of a distribution, including the tail, which cannot be satisfactorily estimated.

The second difficulty is that, even if a curve could be found relating the empirical flash intensity to the mean flash duration mentioned above, to assess whether eqn. (14) was a satisfactory description of it would be no simple matter. The equation contains four unknown constants ν, x, λ and n so that, even if a reasonable fit were obtained, it is doubtful whether it would be particularly meaningful, unless some alternative method could be found of estimating the constants.

7. LONG FLASHES: MOVING AVERAGE PREDICTIONS

It has not proved possible to find an analytic expression for the chance of a long flash being detected on the moving average model. This is not surprising as the moving average process is notoriously intractable (see Naus, 1965). However, it is reasonably easy to simulate the process and since the basic parameter—the period—can be estimated from short flash results, the simulation does not take too long. Only three constants need be varied in the simulation: T, the length of the flash, a the average number of impulses arriving at the accumulator in a period, and n the criterion. Using $P=100$ msec, $M(a, n, T)$ was obtained by simulation where $M(a, n, T)$ is the chance that a flash of duration T, average number of impulses arriving a, and criterion n would be seen. For given values of N and T from Bouman and van der Velden's results, the frequency of seeing Q can be found (see Figure 6a). For these values of Q and T and for a fixed value of the criterion n it is possible to find a value of a such that

$$Q=M(a, n, T).$$

Let $M_n^{-1}(Q, T)$ be the value of a which satisfies this equation for the previously fixed values of Q and T. Now from the axioms, $a=(\nu I+x)P$ where ν, x, P are constant. Therefore,

$$M_n^{-1}(Q, T)=(\nu I+x)P.$$

The value of I for particular values of Q and T is given by the data and therefore this equation can be tested for each point on Bouman and van der Velden's frequency-of-seeing curves for any given value of n. Figure 7 shows the plot of $M_n^{-1}(Q, T)$ against I for 2, 4, 6, 10 and 15 using Bouman and van der Velden's data for the 1 sec and 5 sec flashes and the 60 per cent points for the $\frac{1}{2}$ sec and 2 sec flashes. It can be seen that a reasonable linear fit is obtained, especially for $n=2$.

A second check on the model can be obtained by comparing the obtained straight line with deductions from very short flash experiments. Define the tail of the lag distribution as the area not covered by the central 100 msec. If the lag distribution has negligible tails, all the impulses from a very short flash will contribute together to the moving average; there will not be time for the

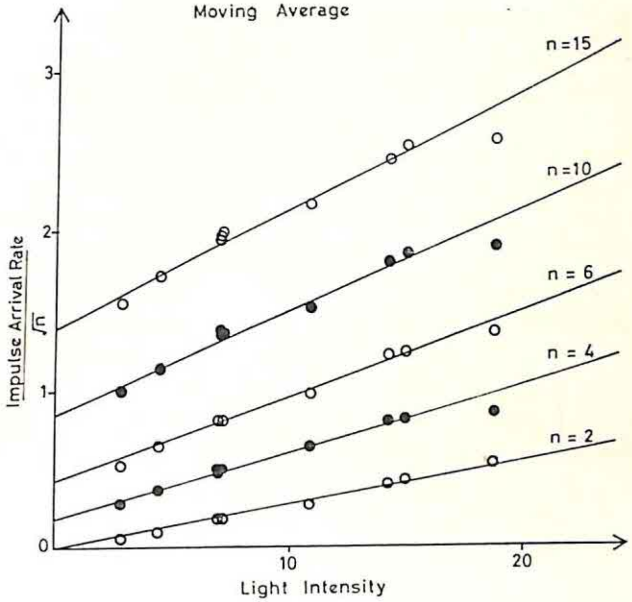


FIGURE 7. Straight line predictions from the moving average model for different values of the threshold, n, fitted by eye to Bouman and van der Velden's long flash data.

first of the impulses to have been in the moving average for more than 100 msec before the last impulse arrives. If a is the average number of impulses arriving at the accumulator from the flash added to the average number of noise impulses in a period and $P_a(n)$ the probability of its being seen with n the criterion, then:

$$P_a(n) = \sum_{y=n}^{\infty} a^y \, e^{-a}/y!\,.$$

Therefore
$$dP/da = e^{-a}\, a^{n-1}/(n-1)!\,.$$

So
$$P_a(n) = \int_0^a e^{-u}\, u^{n-1}/(n-1)!\, du$$
$$= \Gamma_a(n)/\Gamma(n),$$

where $\Gamma_a(n)$ is the incomplete Gamma Function.

The incomplete Gamma Function has been tabulated by Pearson (1922) in the form $G(s_1, s_2)$ where $s_1 = a/\sqrt{n}$ and $s_2 = n-1$. For any value of Q and fixed n, a value of a can be found such that $Q = P_a(n)$. Let $G^{-1}(Q, n-1)$ be this value of a. Consider first $G^{-1}(0{\cdot}6,\ n-1)$. From this, using the values of ν and x obtained from the slope and intercept of the long flash line, the energy of the flash giving the 60 per cent threshold for a very short flash can be estimated from $a = \nu N + xP$. For $n = 2, 4, 6, 10, 15$ this is found to be 52, 46, 38, 36 and 35 quanta respectively. The empirical value obtained by Bouman and van der Velden for an 8 msec flash was 31 quanta. So the prediction for the 60 per cent threshold is fair for the higher values of n but poor for the lower values.

Instead of considering just the 60 per cent threshold it is possible to consider the frequency of seeing curve for very short flashes. Figure 8 shows $G^{-1}(Q, n-1)$ and $M_n^{-1}(Q, T)$ both plotted against I for $n=10$. Since both lines should have intercept x and slope ν they should be the same. They are obviously not. Thus the moving average model for higher values of n is not satisfactory.

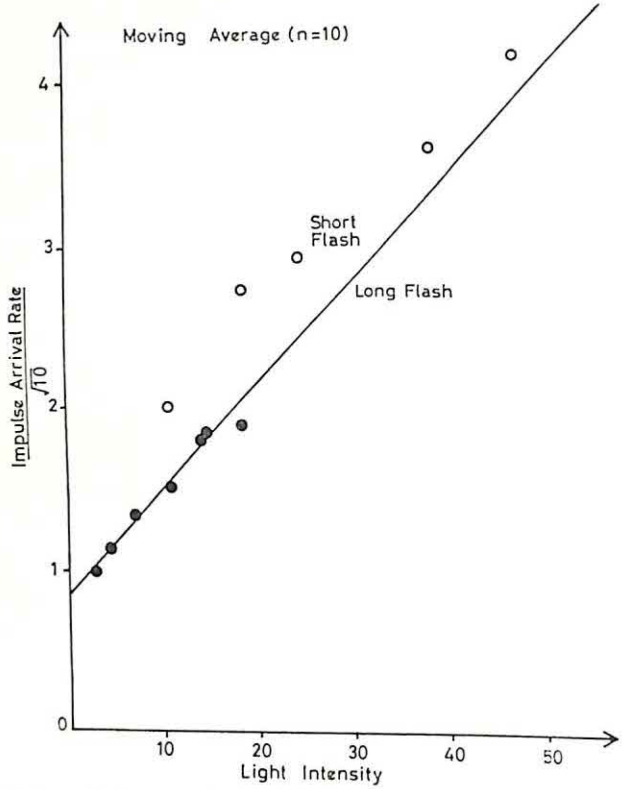


FIGURE 8. Impulse arrival rates derived from Bouman and van der Velden's long and short flash frequency of seeing curves using the moving average model with $n=10$. The short flash points should lie on the straight line which has been drawn by eye through the long flash points.

For the low values of n, say $n=2$, the difficulty for the model is that the threshold energy contained in a short flash is much lower than would be predicted from the long flash results. If the opposite had been true it could be explained by assuming that the tails of the lag distribution could not be neglected, but there is no obvious explanation of the actual result. Altering the value of P is not satisfactory since then the position of the kink at $T=100$ msec becomes impossible to explain. (It would be necessary to assume a period of approximately 35 msec in order to obtain a fit.) One possibility is that the Howarth–Treisman effect is operating (see Section 3) and the criterion is lower for short flashes than it is for long flashes, but this would be very much an *ad hoc* explanation. Also on this explanation one would expect the long flash line to be less steep than the short flash line in Figure 8. It therefore appears that the moving average model cannot provide a satisfactory explanation of long flash results.

8. LONG FLASHES: PERCEPTUAL MOMENT MODEL (POISSON METHOD)

For the moving average model, the probability of a long flash being seen was used to estimate the average rate at which impulses arrive at the accumulator and this should be linearly related to the intensity of the flash. In addition, results from long flash experiments could be compared with results from short flash experiments. A similar approach can be used with the perceptual moment model and proves much simpler since analytic methods can be used throughout.

On the perceptual moment model a long flash has an equal and independent chance of being detected in any moment during its time course. If Q, T, a are defined as in the previous section and $P_a(n)$ is the probability that a flash producing average input a will be seen in any moment, then

$$1-Q=[1-P_a(n)]^{T/D}.$$

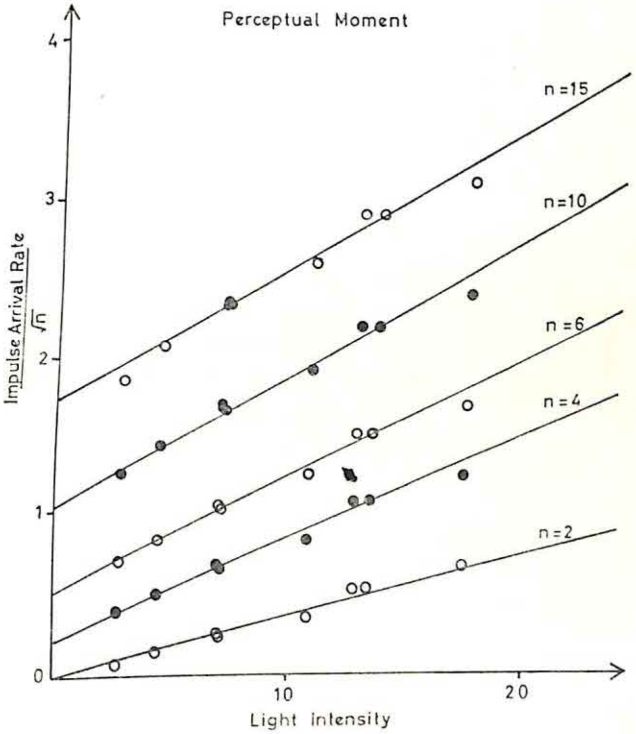


FIGURE 9. Straight line predictions from the perceptual moment model for different values of the threshold, n, fitted by eye to Bouman and van der Velden's long flash data.

However, as before,

$$P_a(n) = \Gamma_a(n)/\Gamma(n) = G(a/\surd(n), n-1).$$

Therefore $$G(a/\surd(n), n-1) = 1-(1-Q)^{D/T}.$$

Let the value of a satisfying the equation be $G^{-1}(Q, n-1, T)$, then we should have

$$\surd(n) G^{-1}(Q, n-1, T) = (vI+x)D.$$

Using the values of Q, T, I used for fitting the moving average model, the existence of the linear fit between I and $G^{-1}(\ )$ can be checked. Figure 9 shows the relation between $a/\surd n$ and I for $n=2$, 4, 6, 10 and 15. $a/\surd n$ is used instead of a as the ordinate because the standard deviation of $G^{-1}(\ )$ is approximately proportional to $\surd n$. Reasonable straight line fits are found for all values of n.

As in the previous section, a second check on the model can be obtained by considering long flash frequency-of-seeing curves in conjunction with short

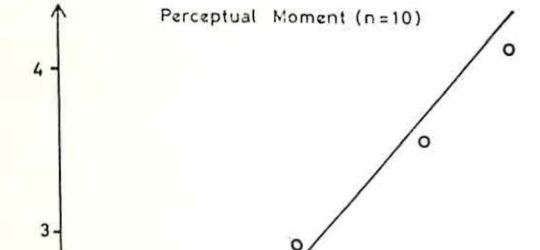


FIGURE 10. Impulse arrival rate derived from Bouman and van der Velden's long and short flash frequency-of-seeing curves using the perceptual moment model with $n=10$. The short flash points should lie on the straight line which has been drawn by eye through the long flash points. Short flash results are again indicated by open circles.

flash curves. However, the assumption that all the impulses from a very short flash summate in one moment is not so reasonable as for the moving average model because, even if there is a small lag variance on some occasions, some of the impulses from the flash will arrive in one moment and the rest in the next. In the following section, evidence will be presented that this effect is not large. Even if it is large, the energy threshold predicted from the long flash results will provide a lower bound to the possible values of the energy. That is to say, the theoretical value cannot be greater than the empirical value. For $n=2$, 4, 6, 10 and 15, the predicted energy for the 60 per cent threshold is 40, 33, 31, 31 and 30 quanta respectively, compared with the empirical value of 31 quanta. For low values of the criterion, the estimate is too high so that it cannot be explained using the idea of lag variance. For higher values, the fit is good. Also, the lines derived from short flash frequency-of-seeing curves are reasonably similar to the lines obtained from long flash curves. An example, for $n=10$, is shown in Figure 10. The lines are much more similar than they were for the moving average model. Therefore the perceptual moment model appears to fit reasonably well for a criterion of not less than about 4 impulses per moment.

9. LONG FLASHES: PERCEPTUAL MOMENT MODEL (BRINDLEY'S METHOD)

Whilst the previous methods enabled predictions to be obtained for small area flashes, the axiom system cannot deal with larger areas where more than one ganglion field is involved. Moreover, since there is strong evidence that ganglion cells have overlapping fields and are therefore not independent (Hartline, 1940; Brindley, 1960 b), simple elaborations of the previous method do not seem possible.

However, Brindley (1960 a) proved a relation which enables generalizations to be made about the effect of duration for large area flashes. Moreover, with this method, no assumption need be made about the relation between a and I, previously assumed linear. He showed that if the I/T curve is of the form $TI^m=c$ for a particular frequency of seeing $Q=Q_0$, then the frequency-of-seeing curves for long flashes, assuming no temporal summation between moments, must be of the form $Q=1-\exp(-kI^m)$ where $k=T/c[\ln 1/(1-Q_0)]$. This is proved by deducing what the frequency-of-seeing curve for one moment must be and hence obtaining the curve for many independent moments.

Brindley (1960 b), also using the results of Bouman and van der Velden, showed that the results for long flashes of small area where $TI^2=c$ for 60 per cent frequency of seeing, produced frequency-of-seeing curves of the predicted form. Brindley's approach can be extended to large area flashes provided their I/T curves have an appropriate form.

From Brindley (1960 a) we should have

$$Q=1-\exp(-kI^m).$$

Therefore

$$\log \ln (1-Q)^{-1}=m \log I+\log k. \tag{15}$$

But for $0{\cdot}3 < Q < 0{\cdot}9$ the following relation holds to a close approximation (see Shallice, 1965):

$$Q = \tfrac{3}{4}\log \ln (1 - Q)^{-1} + 0{\cdot}59.$$

Therefore from eqn. (15) we should have:

$$Q = \tfrac{3}{4}m \, . \log I + \tfrac{3}{4}\log k + 0{\cdot}59. \tag{16}$$

Hence frequency-of-seeing curves should be linear with $\log I$ in the range $0{\cdot}3 < Q < 0{\cdot}9$. This is the case for all areas of flash (see Figure 6). In addition, the slopes of the $(Q, \log I)$ lines can be measured and compared with the shape as predicted from the I/T curves.

If the N/T curve is of the form $NT^k = C$, as it is in all cases to be considered, then it has the equation:

$$\log N = \tan \theta \log T + c_1. \tag{17}$$

Hence $$TI^{(1-\tan \theta)^{-1}} = \exp (c_1).$$

Let the Q/N curve have equation:

$$Q = \tan \phi \log N + c_2.$$

Then from eqns. (16) and (17), $\tan \phi = \tfrac{3}{4}m = \tfrac{3}{4}(1 - \tan \theta)^{-1}$. Therefore

$$\cot \phi = 4/3(1 - \tan \theta).$$

Since the N/T curve has probably less error variance than the individual N/Q curves which are only frequency-of-seeing curves (see Figure 5), θ will be used to estimate ϕ which can be measured directly from Figure 6 for both 1 sec and 5 sec results. A comparison of estimated and empirical values of ϕ is shown in Table 1.

TABLE 1. A COMPARISON OF THE ESTIMATED AND EMPIRICAL VALUES OF ϕ

The tangent of ϕ is the slope of the straight part of the curves in Figure 6, using Brindley's method

Area	Empirical θ	Estimated m	Estimated ϕ	Emp. ϕ (1 sec)	Emp. ϕ (5 sec)
4′	25°	1·9	54°	52°	54°
38′	29°	2·2	59°	60°	65°
213′	36°	3·8	71°	69°	67°
1390′	38°	4·8	74°	75°	77°

The fit can be seen to be good except for the 38′, 5 sec curve and possibly for the 213′, 5 sec curve and is certainly within the margin of error for frequency-of-seeing curves.

Finally, it is possible to estimate what the average energy arriving within a single moment must be to give a 60 per cent chance of detection. This can be compared with the 60 per cent point for very short flashes to see whether the input from a very short flash arrives all in one moment or is spread over a number of moments—an issue that was important in previous sections. Let $P(N')$ be the chance of detecting the input from a flash of energy N' all of whose input arrives at the accumulator in one moment. Therefore

$$(1-P)=(1-Q)^{D/T}=\exp\left[-(D/c)\{\ln(1-Q_0)^{-1}\}I^n\right],$$

so that $$\log \ln(1-P)=m\log N'-(m-1)\log D+\log k.$$

Comparing with eqn. (15), we have, when $P=Q$:

$$\log N-\log N'=\frac{m-1}{m}(\log T-\log D).$$

But from the N/T curve we know that, for large T:

$$\log N=\frac{m-1}{m}\log T+c_1.$$

Hence $$\log N'=\left(1-\frac{1}{m}\right)\log D+c_1.$$

If it is assumed that all the impulses from a very short flash arrive at the accumulator within one moment then the value of D can be read off from Figure 5. It is where the asymptote of the N/T curve has energy equal to that of the shortest flash. From rough measurements from this figure, the following estimates of D were obtained for spots of diameter 4′, 8′, 16′, 38′, 76′, 213′, 663′ and 1390′: 81, 86, 97, 132, 139, 86, 107 and 94 msec respectively. These estimates agree reasonably well with the value $D=100$ msec used throughout. Thus the idea used earlier that lag variance may not greatly influence the predictions receives tentative support.

10. PERIPHERAL THEORIES

Treisman (1966) has shown quite convincingly that the response in an absolute threshold situation is determined centrally. However, it is conceivable that temporal summation could take place at a more peripheral stage of the visual system than response determination. Yet, both the plausible places for temporal summation to occur—in the rods and in the retinal ganglion cell—seem to have drawbacks.

If temporal summation were to occur in the rods, this would mean that the output of the rods would be the same for two flashes of the same energy provided that both lasted for less than the critical duration. However, it has been shown, using statistical arguments, that the perception of a flash at absolute threshold depends normally on a number of rods each absorbing a single photon, provided that the flash has diameter greater than about 10′ (Hecht, Shlaer and Pirenne, 1942; Barlow, 1956; Brindley, 1960 b). If Bloch's law holds for this area of flash, it could not be explained by summation within individual rods. In fact in the first experiment where the effect of area on Bloch's law was investigated, Graham and Margaria (1935) found that whilst temporal summation held up to 100 msec for 2′ diameter flashes, it only held up to about 30 msec for 16′ diameter flashes. However, later experimenters have found that it held almost as well for larger area flashes (e.g. 1° diameter) as it did for flashes of the order of 2′ to 4′ (Bouman and van der Velden, 1947; Barlow, 1958; Baumgardt and Hillman, 1961). Baumgardt and Hillman have produced evidence that Graham and Margaria's results may be distorted by the use of uncalibrated wedges and that such an error is consistent with the results. In any case the weight of evidence suggests that temporal summation occurs at a higher level than the rod.

That temporal summation occurs at the ganglion cell seems at least equally likely, especially as it is known that Ricco's law of complete spatial summation holds for fields of about the area served by a single ganglion cell (Bartlett, 1965). For a larger area, spatial summation seems most likely to be probability summation (Bouman and van den Brink, 1952). This would mean that the brain is not combining the outputs of more than one ganglion cell from an eye in order to make a decision. This idea receives some support at least for the lateral geniculate nucleus where Bishop, Levick and Williams (1964) found that a third order neuron normally receives excitatory input from only one ganglion cell.

How could the ganglion cell operate so as to be the sole locus of temporal summation? Assuming that it operates in the normal fashion for a neuron, then each impulse arriving from a bipolar triggers a burst of transmitter substance across the synaptic gap which in turn produces an increase in excitatory postsynaptic potential (EPSP) in the ganglion cell (see Eccles, 1964). The EPSP decays exponentially until the next burst of transmitter substance arrives. If just after a burst of transmitter substance arrives the EPSP threshold is reached, the cell fires either once or, possibly if the EPSP decay is slow, more than once.

Now if the EPSP builds up to its threshold value more than once during the time the flash is acting, then temporal summation of the impulses produced must occur at a higher level. Hence the only possible mechanism, if temporal summation is to be solely at the retinal level, seems to be that all that is necessary for perception of the flash is for the EPSP to reach its threshold once, and this is sufficient to produce a relatively unambiguous stream of pulses up to the optic nerve since a single impulse would be undetectable in the neural noise which Kuffler, Fitzhugh and Barlow (1957) have demonstrated in such cells. For this to be the case, the EPSP would have to decline very slowly so that more than one impulse could be produced. However, while cells with slow EPSP decays exist, e.g. Renshaw cells, they are rare (see Eccles, 1964).

In addition, Ten Hoopen (1966) has shown that an immigration–death process in which the approximation is used that the EPSP takes only integral values instead of being continuous, provides a satisfactory model of nerve fibre firings. The rise in the EPSP produced by an impulse arriving along a synapsing fibre is taken as the unit. This model is mathematically identical to the GMA so the predictions made from it would be the same as those made by the GMA. They have been shown to be rather unsatisfactory in Section 4. Thus the retinal ganglion cells do not seem likely to be the sole locus of temporal summation.

More conclusive than these arguments, which depend on extrapolations from present-day physiological knowledge, is some psychological evidence. Matin (1962) found that if two flashes are presented to corresponding points on the retina the chance of detection is higher for inter-flash intervals up to 100 msec than would be predicted if probability summation alone was operating. Whilst, as Matin points out, some of the preceding work on this problem had failed to detect summation, this seems strong evidence that some temporal summation is occurring centrally.

11. Discussion

The arguments of the preceding section make it implausible that temporal summation can be explained entirely peripherally. However, the axioms of Section 2 are not the only hypotheses consistent with the arguments of the preceding section. Three other such hypotheses are those of Baumgardt (1953), Treisman (1964 b, 1966), and Blackwell (1963).

Baumgardt suggested that a flash is not detected if several quanta are absorbed in a sensitive unit, but only if a number of sensitive units each absorb, nearly simultaneously, two or more quanta. This could easily be fitted in with the arguments of the preceding section by assuming that the EPSP threshold of a ganglion cell is between one and two times the rise in EPSP produced by the effect of one quantum absorbed in the ganglion's field. The integration of two-quanta coincidence would take place more centrally. It seems very difficult to work out what the predictions of this theory would be, taking into account the complications discussed in this paper. Baumgardt only discussed the predictions of his theory with respect to spatial summation, so that as far as temporal summation is concerned his theory remains at present uncorroborated, unlike, say, the perceptual moment model.

Treisman suggests that the effective photons absorbed trigger not one impulse in the ganglion cells but a train of them. However, Treisman, in the development of his theory in order to fit Barlow's (1957) findings, assumes that at low light intensities the variability in the pulse trains produced by each effective photon is negligible. So the extra complication of assuming variability in the pulse trains would not affect the predictions made in the present paper.

Blackwell's theory, as discussed in Section 1, is conceptually almost totally distinct from all other work in the area with its emphasis on the lag distributions rather than summation. It provides a good fit to his data for short flashes, but in order to fit the long flash data he has to introduce a second assumption which is equivalent to a moment model. This is because he phrases his theory in terms of a continuous variable instead of using a stochastic formulation. Since the moment model alone will explain the results, there seems no need for his rather stringent assumptions about the form of the lag distribution. Whether, with these assumptions alone translated into stochastic terms, his theory could explain the long flash results is difficult to ascertain, as the theory is difficult to express mathematically. Hence, despite its conceptual elegance and simplicity, it does not seem a satisfactory theory at present.

Returning to the Change Detector models of this paper, some of the tests of the individual models are critically dependent on the axioms used, others are relatively independent of them. Three sets of predictions are not critically dependent on the precise set of peripheral axioms used: those of Sections 5 and 9 concerning the perceptual moment model and the prediction used in Sections 7 and 8 that the long and short flash results should lie on the same curve of intensity plotted against impulse arrival rate. The predictions in Section 5 depend upon an assumption about the relation between the chance of detection in a moment and the stimulus energy giving rise to the input to the accumulator in that moment. This relation is assumed to be given by the empirical relation between frequency of seeing and intensity for a very short flash. This implies that the lag distribution has zero variance, but the analysis in Appendix 2, where this condition is relaxed, is also not critically dependent on the axioms. In Section 5 it is shown that the perceptual moment model and the moving average models are about equally good fits for the short flash data provided complete temporal summation is not found. In Section 9 no assumption is made other than that the Change Detector operates on the perceptual moment model. It is found to fit the data well.

In Sections 7 and 8 most of the predictions are critically dependent on the axioms and here the perceptual moment model is found to be superior to the moving average model in its explanation of the long flash data. Finally, in Section 4 two further assumptions are made that the lag distribution is an exponential distribution and that the stochastic model can be approximated by a continuous one; and it is found that the GMA cannot explain the experimental results satisfactorily.

The perceptual moment model is the only one supported and that by predictions not dependent on other assumptions. The moving average model and the GMA have not been refuted since they have not been tested when isolated from other assumptions, but no support has been found for them. These theoretical deductions are dependent on two conclusions about experimental results both of which need further testing:

(1) That complete summation is not obtained at absolute threshold for up to 100 msec flashes and in particular that Baumgardt and Hillman's results are subject to some artefact.

(2) That Bouman and van der Velden's findings for long flash thresholds are not subject to any artefact.

APPENDIX 1

THAT THE GMA DOES NOT FIT SHORT FLASH CURVES

It was shown that the GMA predicts that the N/T curve for short flashes should obey the following equation:

$$N = EbT\frac{(1-\mathrm{e}^{-bT})^{c/(b-c)}}{(1-\mathrm{e}^{-cT})^{b/(b-c)}}, \text{ where } E,\ b,\ c \text{ are constant.}$$

Let $b = rc$, therefore

$$\begin{aligned}\frac{N(2T)}{N(T)} &= 2\,\frac{(1-\mathrm{e}^{-2rcT})^{1/(r-1)}}{(1-\mathrm{e}^{-2cT})^{r/(r-1)}} \times \frac{(1-\mathrm{e}^{-cT})^{r/(r-1)}}{(1-\mathrm{e}^{-rcT})^{1/(r-1)}}\\ &= 2\,\frac{(1+\mathrm{e}^{-rcT})^{1/(r-1)}}{(1+\mathrm{e}^{-cT})^{r/(r-1)}}\\ &= 2\,\frac{(1+a^{r})^{1/(r-1)}}{(1+a)^{r/(r-1)}}, \text{ where } a = \mathrm{e}^{-cT}.\end{aligned}$$

Empirically, this fraction remains very close to unity and then increases rapidly before becoming relatively stable. Is this theoretically possible?

Consider $$\begin{aligned}g(T) &= \left[\frac{N(4T)}{N(2T)} - 1\right]\left[\frac{N(2)T}{N(T)} - 1\right]^{-1}\\ &= \left[2\,\frac{(1+a^{2r})^{1/(r-1)}}{(1+a^{2})^{r/(r-1)}} - 1\right]\left[2\,\frac{(1+a^{r})^{1/(r-1)}}{(1+a)^{r/(r-1)}} - 1\right]^{-1}.\end{aligned}$$

It can be shown by the method of approximations that, for any value of r, as $a \to 1$, $g(T) \to 4$ and as $a \to 0$, $g(T) \to 1$. By computer exploration of the function it was found that $dg/da > 0$ for $0 < a < 1$, so that for $0 < a < 1$, $g(T) < 4$. The data which has the least sharp kink and is therefore most likely to fit is that of Bouman and van der Velden. Using their data and putting $T = 50$ msec, we have $g(T) = 4{\cdot}5$. By itself this would not necessarily mean that their data could not be fitted by the model, since the supra-maximum value could be attributed to slight errors in the values used. This is plausible since $g(T)$ is related to d^2N/dT^2 and is not therefore likely to be very accurate.

However, for the empirical data, we find $g(2T) \simeq 1$ when $T = 50$ msec: that is, the N/T curve is linear in the region 100–400 msec. Yet from the model, if $g(T)$ is large, $g(2T)$ must also be large. Table 2 shows for a given value of $g(T)$, what is the lowest value that $g(2T)$ can take for any possible value of a. This makes an explanation in terms of slight errors less likely. No doubt by slightly increasing its complexity the GMA could be made to fit the data, but it is not now an obvious choice for a model of brightness thresholds.

TABLE 2. THE RESULTS OF A COMPUTER EXPLORATION OF $g(2T)$ FOR PARTICULAR VALUES OF $g(T)$

$g(T)$	3·75	3·5	3·25	3·0	2·5
min $[g(2T)]$	3·10	2·48	2·05	1·75	1·37

APPENDIX 2

THE EFFECT ON THE PERCEPTUAL MOMENT MODEL OF INTRODUCING LAG SPREAD

Consider two perceptual moment models—M, the one with zero lag variance considered earlier, and also M′ which includes lag spread. On model M′ for a flash of energy N/ν lasting from $-\tau$ to $-\tau+h$, let the chance of an impulse produced by the flash arriving at the accumulator between t and $t+\delta t$ be:

$$\begin{cases} Nf(t)\,\delta t & \text{if } 0 \leqslant t < L (L < D), \\ 0 & \text{otherwise,} \end{cases} \tag{18}$$

where τ is the minimum transmission lag and $f(t)$ is a non-negative function determining the lag spread and satisfying $\int_0^L f(t)\,dt = 1$. Hence the input to the M′ accumulator from the flash will be the same as the input to the M accumulator (with zero transmission lag) from a flash whose intensity is described by eqn. (18). Hence, since the system is assumed linear, the effect of lag spread may be investigated by using the zero lag variance model but with a changed stimulus input.

Consider four possible inputs to model M. Two represent the input to model M itself—a very short flash (F_1) of energy N and a flash of (F_2) of length D and energy N'. The other two represent the corresponding transformed inputs to model M′—a flash (F_3) described by eqn. (18) and a flash (F_4) described by the following equations:

$$I(t) = \begin{cases} N' \int_0^t f(t)dt/D & 0 \leqslant t \leqslant L; \\ N'/D & L \leqslant t \leqslant D; \\ N' \int_t^L f(t)dt/D & D \leqslant t \leqslant D+L; \\ 0 & \text{otherwise.} \end{cases}$$

Let t be the time from when the flash starts to the next moment boundary and let $J_i(t)$ be the most energy from the flash F_i which occurs in any one moment. Throughout this section it will be assumed that the flash is only detected in the maximum energy moment and also that $\tau = 0$. This second assumption is only made to ease the exposition and does not affect the conclusions. If $t > L$, $J_1(t) = J_3(t)$ since all the energy arrives in one moment. If

$$t < L,\ J_1(t) - J_3(t) = \min\left[\int_0^t Nf(t)\,dt,\ \int_t^L Nf(t)\,dt\right] > 0. \tag{19}$$

Define B by the equation

$$L - B = \int_0^L \int_0^t f(x)\,dx\,dt.$$

Then if $t > L$,

$$J_2(t-B) = J_4(t). \tag{20}$$

If $t<B$,

$$0<J_2(t-B)-J_4(t)<2tN'\int_0^t f(t)\,dt/D<2tN'[J_1(t)-J_3(t)]/ND. \tag{21}$$

Similarly, if $B<t<L$,

$$0<J_2(t-B)-J_4(t)<2(L-t)N'[J_1(t)-J_3(t)]/ND. \tag{22}$$

Consider

$$Q\equiv[Q_1(N)-Q_3(N)]-[Q_2(N')-Q_4(N')],$$

where $Q_i(N)$ is the probability of seeing flash F_i. Considering variations in the moment boundary:

$$Q=\int_0^D\{[Q_1(J,t)-Q_3(J,t)]-[Q_2(J,t-B)-Q_4(t)]\}dt.$$

Since $J(t)=J(D+t)$, inserting $t-B$ instead of t rearranges the contents of the integral without affecting its sum. Using the mean value theorem since it may reasonably be assumed that Q is a continuous function of J:

$$Q=\int_0^D\left\{[J_1(t)-J_3(t)]\left(\frac{dQ}{dJ}\right)_{J=J^*}-[J_2(t-B)-J_4(t)]\left(\frac{dQ}{dJ}\right)_{J=J^{**}}\right\}dt,$$

where $J^*(t)$, $J^{**}(t)$ lie within the range:

$$J_3(t)\leqslant J^*(t)\leqslant J_1(t);\quad J_4(t)\leqslant J^{**}(t)\leqslant J_2(t-B)\quad\text{for all } t.$$

Therefore

$$Q>\int_0^D\left\{[J_1(t)-J_3(t)]\left(\frac{dQ}{dJ}\right)_{J=J'}-[J_2(t-B)-J_4(t)]\left(\frac{dQ}{dJ}\right)_{J=J''}\right\}dt,$$

where

$$\left(\frac{dQ}{dJ}\right)_{J=J'}=\min\left(\frac{dQ}{dJ}\right)\text{ for } N/2\leqslant J\leqslant N;$$

$$\left(\frac{dQ}{dJ}\right)_{J=J''}=\max\left(\frac{dQ}{dJ}\right)\text{ for } N'/2\leqslant J\leqslant N'.$$

Let

$$\left(\frac{dQ}{dJ}\right)_{J=J''}\Big/\left(\frac{dQ}{dJ}\right)_{J=J'}=\alpha.$$

Since Q is a monotonically increasing function of J, $\alpha>0$, $\left(\frac{dQ}{dJ}\right)_{J=J'}>0$. Then

$$Q>\left(\frac{dQ}{dJ}\right)_{J=J'}(L/D)\int_0^L\{[J_1(t)-J_3(t)]-\alpha[J_2(t-B)-J_4(t)]\}\,dt,$$

since if $t>L$, $J_1(t)=J_3(t)$ and $J_2(t-B)=J_4(t)$ from eqns. (19) and (20). Therefore using eqns. (21) and (22),

$$Q>\left(\frac{dQ}{dJ}\right)_{J=J'}(L/D)\int_0^B\left\{[J_1(t)-J_3(t)]\left[1-\frac{\alpha 2tN'}{ND}\right]dt\right.$$
$$\left.+\int_B^L[J_1(t)-J_3(t)]\left[1-\frac{2(L-t)\alpha N'}{ND}\right]\right\}dt.$$

This last expression is necessarily greater than zero if $ND>2\alpha BN'$ and $ND>2\alpha N'(L-B)$. Assuming that L and B are such that $Q_3(N)$ is at the desired frequency of seeing, let $\log N' = \log N + 0{\cdot}13$ so that $Q_1(N) = Q_2(N')$ from Section 5. Since $Q > 0$, $Q_4(N') > Q_3(N)$.

Hence if N'' is the intensity at which $Q_4(N'') = Q_3(N)$ then

$$\log N'' < \log N' < \log N + 0{\cdot}13$$

so that the difference between the threshold energies for $t = 0$ and $t = D$ is decreased by introducing a lag.

The proof holds if $ND > 2\alpha BN'$ and $ND > 2\alpha N'(L - B)$, i.e. if $B < ND/2\alpha N'$ and $L - B < ND/2\alpha N'$, both of which depend on α. α can be estimated from the empirical data, e.g. Barlow (1956), but because it depends on the slope of the frequency-of-seeing curve only a rough estimate can be made. For the 60 per cent point α is quite close to unity. If one assumes that the frequency-of-seeing curves are Poisson sums, then a more accurate estimate of α can be made, but it depends upon the criterion n. Considering the 60 per cent frequency of seeing for $n = 2, 4, 6, 10, 15$, $\alpha = 1{\cdot}39, 1{\cdot}21, 1{\cdot}16, 1{\cdot}11$ and $1{\cdot}10$. Taking $\alpha = 1{\cdot}2$ as a maximum (since $n = 2$ seemed unlikely from Section 8), we have that the proof holds if $B < 0{\cdot}33\ D$ and $L - B < 0{\cdot}33\ D$.

However, examination of the proof shows that it contains a series of inequalities each of which is normally quite pronounced so that the result probably holds for a much wider range of lags.

Acknowledgements

I should like to thank Dr. Donald Laming for his valuable comments on the first draft of this paper and Dr. Andrew Gregory and Professor R. J. Audley for the stimulus of conversations with them on various aspects of the research. The research was mainly carried out while I received a Science Research Council Research Fellowship.

References

Allport, A. (1966). *Phenomenal simultaneity and the perceptual moment hypothesis.* Unpublished Ph.D. thesis, University of Cambridge.

Audley, R. J. (1963). Decision models in reaction time. In: Psychophysics and the ideal observer. *Proc. 17th Int. Congr. Psychol.*, pp. 236–237. Amsterdam: North-Holland.

Barlow, H. B. (1956). Retinal noise and absolute threshold. *J. opt. Soc. Amer.* **46,** 634–639.

Barlow, H. B. (1957). Increment thresholds at low intensities considered as signal-noise discriminations. *J. Physiol.* **136,** 469–488.

Barlow, H. B. (1958). Temporal and spatial summation in human vision at different background intensities. *J. Physiol.* **141,** 337–350.

Barlow, H. B., Fitzhugh, R. and Kuffler, S. W. (1957). Change of organization in the receptive fields in the cat's retina during dark adaptation. *J. Physiol.* **137,** 338–354.

Bartlett, N. R. (1965). Thresholds as dependent on some energy relations and characteristics of the subject. In C. H. Graham (ed.), *Vision and Visual Perception.* New York: Wiley.

Baumgardt, E. (1953). Seuils visuels et quanta de lumière. Précisions. *Année psychol.* 53, 431–441.

BAUMGARDT, E. and HILLMAN, B. (1961). Duration and size as determinants of peripheral retinal response. *J. opt. Soc. Amer.* **51,** 340–344.

BAUMGARDT, E. and SMITH, S. W. (1965). Facilitation effect of background light on target detection: a test of theories of absolute threshold. *Vision Res.* **5,** 299–312.

BAUMGARDT, E. and SMITH, S. W. (1966). Reply to Nachmias concerning interpretation of a study to test theories of absolute threshold. *Vision Res.* **6,** 117–118.

BISHOP, P. O., LEVICK, W. R. and WILLIAMS, W. O. (1964). Statistical analysis of the dark discharge of lateral geniculate neurons. *J. Physiol.* **170,** 598–612.

BLACKWELL, H. R. (1963). Neural theories of simple visual discriminations. *J. opt. Soc. Amer.* **53,** 129–160.

BOUMAN, M. A. and VAN DEN BRINK, G. (1952). On the integrative capacity in time and space of the human peripheral retina. *J. opt. Soc. Amer.* **42,** 617–620.

BOUMAN, M. A. and VAN DER VELDEN, H. A. (1947). The two-quanta explanation of the dependence of the threshold values and visual acuity on the visual angle and the time of observation. *J. opt. Soc. Amer.* **37,** 908–919.

BOYNTON, R. M. and SIEGFRIED, J. B. (1962). Psycho-physical estimates of on-responses to brief light flashes. *J. opt. Soc. Amer.* **52,** 720–721.

BRINDLEY, G. S. (1960 a). Two more visual theorems. *Quart. J. exp. Psychol.* **12,** 110–112.

BRINDLEY, G. S. (1960 b). *Physiology of the Retina and the Visual Pathways.* London: Edward Arnold.

COX, D. R. and MILLER, H. D. (1965). *The Theory of Stochastic Processes.* London: Methuen.

DAVY, E. (1952). The intensity–time relation for multiple flashes of light in the peripheral retina. *J. opt. Soc. Amer.* **42,** 937–941.

ECCLES, J. C. (1964). *The Physiology of Synapses.* Berlin: Springer.

FELLER, W. (1957). *An Introduction to Probability Theory and its Applications.* New York: Wiley.

FITZHUGH, R. (1957). The statistical detection of threshold signals in the retina. *J. gen. Physiol.* **40,** 925–948.

GRAHAM, C. H. and MARGARIA, R. (1935). Area and the intensity–time relation in the peripheral retina. *Amer. J. Physiol.* **113,** 299–305.

GREGORY, A. H. (1963). *The quantum efficiency of human vision.* Unpublished Ph.D. thesis, Hull University.

HARTER, M. R. (1967). Excitability cycles and cortical scanning. *Psychol. Bull.* **68,** 47–58.

HARTLINE, H. K. (1934). Intensity and duration in the excitation of single photoreceptor units. *J. cell. comp. Physiol.* **5,** 229–247.

HARTLINE, H. K. (1940). The receptive fields of optic nerve fibres. *Amer. J. Physiol.* **130,** 690–699.

HECHT, S., SHLAER, S. and PIRENNE, M. (1942). Energy, quanta and vision. *J. gen. Physiol.* **25,** 819–840.

HOWARTH, C. I. and TREISMAN, M. (1958). The effect of warning interval on electric phosphene and auditory thresholds. *Quart. J. exp. Psychol.* **10,** 130–141.

JOHNSON, E. P. (1958). The character of the b-wave in the human electroretinogram. *Archiv. Ophthal.* **60,** 565–589.

KAHNEMAN, D. and NORMAN, J. (1964). The time-intensity relation as a function of the observer's task. *J. exp. Psychol.* **68,** 215–200.

KUFFLER, S. W., FITZHUGH, R. and BARLOW, H. B. (1957). Maintained activity in the cat's retina in light and darkness. *J. gen. Physiol.* **40,** 683–702.

LABERGE, D. (1962). A recruitment theory of simple behaviour. *Psychometrika* **27,** 375–396.

LAMING, D. R. J. (1968). *The Information Theory of Human Choice–Reaction Times.* New York: Academic Press.

LONG, G. E. (1951). The effect of duration of the onset and cessation of a light flash on the intensity–time relation in the peripheral retina. *J. opt. Soc. Amer.* **41,** 743–747.

LUCE, R. D. (1963). Detection and recognition. In *Handbook of Mathematical Psychology*, Vol. 1 (ed. R. D. LUCE, R. R. BUSH, E. GALANTER). New York: Wiley.

McGILL, W. J. (1963). Stochastic latency mechanisms. In *Handbook of Mathematical Psychology*, Vol. 1 (ed. R. D. LUCE, R. R. BUSH, E. GALANTER). New York: Wiley.

MATIN, L. (1962). Binocular summation at the absolute threshold of peripheral vision. *J. opt. Soc. Amer.* **52,** 1276–1286.

MUELLER, C. G. (1954). A quantitative theory of visual excitation for the single photoreceptor. *Proc. Nat. Acad. Sci.* **40,** 853–863.

NACHMIAS, J. (1966). Photosensitization and intrinsic noise in the visual system. *Vision Res.* **6,** 113–115.

NACHMIAS, J. and STEINMAN, R. M. (1963). Study of absolute visual detection by the rating scale method. *J. opt. Soc. Amer.* **53,** 1206–1213.

NAUS, J. I. (1965). Clustering of random points in two dimensions. *Biometrika* **52,** 263–267.

OSGOOD, C. E. (1953). *Method and Theory in Experimental Psychology*. New York: Oxford University Press.

PEARSON, K. (1922). *Tables of the Incomplete Gamma-Function*. London: H.M.S.O.

PIERON, H. (1920). De la variation de l'énergie luminaire en fonction de la durée de l'excitation pour la vision fovéale. *C.R. de l'Acad. Sci.* **170,** 525–528.

PIKE, A. R. (1966). Stochastic models of choice behaviour: response probabilities and latencies of finite Markov chain systems. *Br. J. math. statist. Psychol.* **19,** 15–32.

RAAB, D. H. (1962). Statistical facilitation of simple reaction times. *Trans. N.Y. Acad. Sci.* **24,** 574–590.

RIGGS, L. A., ARMINGTON, J. C. and RATLIFF, F. (1954). Motions of the retinal image during fixation. *J. opt. Soc. Amer.* **44,** 315–321.

ROUSE, R. O. (1952). Colour and the intensity–time relation. *J. opt. Soc. Amer.* **42,** 626–630.

SHALLICE, T. (1964). The detection of change and the perceptual moment hypothesis. *Br. J. statist. Psychol.* **17,** 113–135.

SHALLICE, T. (1965). *The perceptual moment hypothesis and the detection of change.* Unpublished Ph.D. thesis, University of London.

SHALLICE, T. and VICKERS, D. (1964). Theories and experiments on discrimination times. *Ergonomics* **7,** 37–49.

SPERLING, H. G. and JOLLIFFE, C. L. (1965). Intensity–time relation at threshold for spectral stimuli in human vision. *J. opt. Soc. Amer.* **55,** 191–199.

STROUD, J. M. (1955). The fine structure of psychological time. In *Information Theory in Psychology* (ed. H. Quastler). Glencoe, Ill.: Free Press.

TEN HOOPEN, M. (1966). Probabilistic firing of neurons considered as a first passage problem. *Biophys. J.* **6,** 435–451.

TREISMAN, M. (1964 a). The effect of one stimulus on the threshold for another: an application of signal detectability theory. *Br. J. statist. Psychol.* **17,** 15–35.

TREISMAN, M. (1964 b). Noise and Weber's Law: the discrimination of brightness and other dimensions. *Psychol. Rev.* **71,** 314–330.

TREISMAN, M. (1966). A statistical decision model for sensory discrimination which predicts Weber's Law and other sensory laws: some results of a computer simulation. *Percept. and Psychophys.* **1,** 203–230.

VAN DER VELDEN, H. A. (1944). Over het aantal lichtquanta dat nodig is voor een lichtprikkel bij het menselijk oog. *Physica* **11,** 179–189.

ZWISLOCKI, J. (1960). Theory of temporal auditory summation. *J. acoust. Soc. Amer.* **32,** 1046–1060.

Vol. 20
Part 2
163–186

The British Journal of Mathematical and
Statistical Psychology

November
1967

NON-PARAMETRIC TREND TESTS FOR LEARNING DATA

By A. R. Jonckheere and Gordon H. Bower
University College London Stanford University

Many current stochastic models of learning imply stationarity of portions of individual response sequences. Non-parametric tests are presented for detecting overall group trend and homogeneity of individual trends in a sample of protocols. The protocols may differ in length and consist of any responses that take ordered values, such as errors and successes, ratings, or response latencies. The power of the test for overall group trend is discussed in relation to both linear and logistic alternatives.

1. Introduction

Recent developments in mathematical learning theory have revealed the need for statistical tests of trends in response sequences. Such tests are of importance when applying the one-stage or two-stage all-or-none learning models to empirical data (Bower, 1961; Bower and Theios, 1964) since the strongest parameter-free prediction of these models is that there will be no trend over certain portions of the response sequences obtained from individual subjects. Examples of this might be the sequence of trials prior to the last error, or the trials lying between the first success and the last error; and these predictions of stationarity hold whether success–error sequences, or response latencies are considered.

We shall propose a statistical test for the detection of such trends which is powerful, and applicable to all cases known to us. In the following, though we shall focus mainly on short error–success sequences obtained from learning experiments, the results are applicable to sequences of any length consisting of responses that can take ordered values, e.g., latencies, response amplitudes, or confidence judgements.

In learning experiments, the response sequences which are examined for trend have the following peculiarities. First, the performance measure often has only nominal significance, as with success–error sequences, and has to be correlated with a second variable, the trial number, which has only ordinal significance. Second, the composition and length of each protocol are usually random variables, different subjects having differing numbers of successes and errors, and different trial numbers for their last errors. Hence classical regression analyses (e.g., orthogonal polynomials) are inapplicable, even if their restrictive assumptions could be met.

2. THE VINCENTIZED CHI-SQUARE AND ITS PROBLEMS

Of several tests for stationarity of success–error sequences examined by Suppes and Ginsberg (1963), the Vincentized chi-square was recommended as being the least biased by subject and/or trial-selection artifacts and is now in wide use. In this method, each response sequence is divided into B equal, non-overlapping, trial blocks, and the number of successes and errors tabulated in each block. The $2 \times B$ contingency table obtained from each protocol is then summed over all protocols, and the null hypothesis, that the proportion of successes is the same in each Vincentized trial block, is then tested by the chi-square statistic, with $B-1$ degrees of freedom, computed from this overall table.

We shall discuss the deficiencies of this test before proposing an alternative method. They concern both the Vincentizing procedure and the chi-square test. The procedural problems of the Vincentizing method have been discussed elsewhere (Hilgard, 1938); here it will suffice to note that no rational grounds can be provided for deciding what B should be, nor for deciding what to do with any protocol whose trial-length is not divisible by B. Current practices are both arbitrary and often involve discarding data. Further, the rationale of the Vincent method seems doubtful in the light of current models for learning. For it is assumed that all subjects have the same deterministic learning curve, differing only in a rate parameter, and that their curves could be superimposed by uniformly stretching or contracting the trials-scale for each subject. However, in view of the descriptive success of recent learning models, using either subject-controlled operators or Markov chains, this rationale is extremely questionable.

There are also several deficiencies in the use of the chi-square statistic as a test for trend. First, it is applicable only with frequency data, such as success–error sequences, and is inapplicable to metric data, such as latencies. Second, the Suppes and Ginsberg (1963) test examines *group* average results and ignores possible differences in trends between subjects. Thus, if half the subjects increased and half decreased their success probabilities over trials, the null hypothesis would be accepted. Finally, and most importantly, the chi-square is not a specific test for trend at all, since the ordering of the columns in a $2 \times B$ contingency table is immaterial to its chi-square value. When the chi-square test is used, the alternative to the null hypothesis is simply its negation. Thus the test is sensitive to all possible kinds of deviations from the null hypothesis of stationarity and is consequently less powerful in detecting trends than suitably tailored tests.

The basic statistic to be proposed, Kendall's S (Kendall, 1955), answers all the above criticisms. It is a test which is powerful for detecting trends; it is generally applicable, whether the response measure has nominal, ordinal, or metric significance; the homogeneity of individual subjects' trends can be tested as well as the group average trend; and the questionable assumptions of the Vincentizing method are avoided.

3. THE TEST PROPOSED

Notation We suppose that a subject's response on each trial can be assigned to one of $c \geqslant 2$ categories which can be strictly ranked in some *a priori* way by the experimenter. Examples of ranked categories would be success–error ($c=2$), success–error in combination with one of five confidence ratings ($c=10$), or response latencies (e.g., $c=150$ if that is the number of distinct latencies a subject produces in a sequence of trials). The most frequent, and intuitively natural, ranking of the c response categories is in terms of their relative accuracy of performance: correct responses are better than errors, correct with high confidence are better than correct with low confidence, and fast responses are better than slow ones.

Considering the ith protocol in a sample of N, let n_i be the number of trials, and n_{ik} the overall frequency of response category k, where $i=1, \ldots, N$; and $k=1, \ldots, c$. Later, when we refer to a general protocol and wish to suppress the protocol number in that context, we use n for the number of trials and n_k for the number of occurrences of category k. The background theory indicates which trials are to be inspected for trend or the lack of it. We shall suppose that they have been extracted from each protocol, with the earliest trial labelled as trial 1.

The null hypothesis is then that all distinct permutations of the n_i responses are equally likely to occur. The general alternative trend hypothesis is that the lower-numbered categories of response become more probable as trials increase.

Calculation of S The ith protocol can be rewritten as an ordered $c \times n_i$ contingency table with response categories as rows and trial numbers as columns. Each column will contain one entry while the total number of entries in row k will be n_{ik}.

The computation of Kendall's S (see e.g., Hays, 1963, pp. 649–650) for any such ordered contingency table can be carried out by the following algorithm. For each cell of the contingency table in turn, take the total of the frequencies in the cells lying ' south-east ' (below, right) of it, subtract the total of the frequencies in the cells lying ' south-west ' (below, left) of it, and multiply the result by the frequency in the cell considered. Summing the results over all cells yields S. Two numerical examples are given below: the first is typical of individual learning error-success sequences; the second has three possible categories of response.

EXAMPLE 1

	Trials				
	1	2	3	4	5
Error	1			1	
Success		1	1		1

for which $S=1(3-0)+0(2-0)+0(1-1)+1(1-2)=2$.

EXAMPLE 2

		Trials				
		1	2	3	4	5
Response	3	0	1	0	0	0
Categories	2	1	0	1	0	0
	1	0	0	0	1	1

for which $S=1(3-1)+1(2-0)+1(2-0)=6$, the values for the zero cell entries having been omitted. The directional ordering of the response (row) categories is immaterial, so long as one is consistent: one order gives $+S$, the opposite gives the complement or $-S$. We have adopted the convention that an increasing proportion of ' successes ' over trials is reflected by positive S scores.

Equivalent forms of the algorithm can be used directly on the sequences instead of the $c \times n$ matrices. One equivalent form is: assign to each trial in the sequence a score consisting of the frequency of higher-numbered categories which follow it, minus the frequency of lower-numbered categories which follow it, and add these difference scores over trials. Thus, the sequence 21312231 yields an S of $(2-3)+(4-0)+(0-4)+(3-0)+(1-1)+(1-1)+(0-1)$, which equals 1.

For binary (success–error) sequences with n_1 successes and n_2 errors, if we define u to be the number of successes which follow errors (summed over all errors), then

$$S=2u-n_1n_2.$$

For example, the sequence ESEESSSE has $u=4+3+3+0=10$ and $S=2(10)-4{\cdot}4=4$. The factor n_1n_2 represents the maximum possible value of u. If instead one counts errors which follow successes, to get different numbers u' and $S'=2u'-n_1n_2$, then S' is the negative of S. For our illustrative sequence, $u'=6$ and $S'=-4$. Since S' is the negative of S, it follows that $u+u'=n_1n_2$. The calculational procedure decided upon must be used with all protocols.

S is not defined for sequences having occurrences of only one response category (e.g., all successes). Such protocols obviously cannot be used for a test based upon the possible permutations of the response categories. In the following, when N denotes the number of protocols, it is to be understood that these are protocols having $n_in_j \geqslant 0$ for some categories i and j.

Null distribution of S On the null hypothesis, given the marginal frequencies n_{ik}, all trial-wise permutations of these response categories are equally likely. In this case, Kendall (1955) showed that S_i, the value of S for the ith protocol, approaches (with increasing n_{ik}) to a normal distribution, with mean zero and variance given by

$$\sigma_i^2=\frac{1}{9}\left(n_i^3-\sum_{k=1}^{c} n_{ik}^3\right)+\frac{1}{6}\left(n_i^2-\sum_{k=1}^{c} n_{ik}^2\right), \qquad (1)$$

where n_{ik} is the frequency of category k for sequence i, and n_i is the total number of trials in sequence i. For a dichotomy ($c=2$), as in success–error sequences, eqn. (1) reduces to

$$\sigma_i^2=(n_{i_1}n_{i_2})(n_i+1)/3. \tag{2}$$

From the foregoing it follows that $x_i=S_i/\sigma_i$ approaches to a unit normal deviate. Hence x_i would provide a test statistic for the null hypothesis for the ith sequence. Of course, for small n_i one would be concerned with the approximation of x_i to normality. We consider this matter later, since it is not often that one is interested in a significance test for single protocols.

However, before developing our group test, it may be of interest to point out the relation of Kendall's S on a dichotomy (for a single protocol) to the Mann–Whitney U test and the Wilcoxon test (cf. Siegel, 1956). The Mann–Whitney U statistic is equal to our u (or u', whichever is smaller), and hence U is a simple transform of S. The Wilcoxon T statistic is a sum of ranks (or trial numbers of successes, say) which is related to S according to

$$T=\frac{S}{2}+\frac{n_1}{2}(n+1). \tag{3}$$

While the Mann–Whitney and Wilcoxon tests are usually viewed as tests for whether two groups differ, for our dichotomous case the two ' groups ' being compared are the successes and errors in a protocol, which are ranked by their trial numbers of occurrence. If tests are desired for single short sequences, where the normality approximation of S is in doubt, then the Mann–Whitney or Wilcoxon statistics should be used (instead of S) since their exact probability levels are already extensively tabled for small n_1 and n_2 values (cf. Siegel, 1956). The decision, of course, would be equivalent to that obtained from calculating exact significance values of S, since U and T are merely linear functions of S.

Test for group trend Let N denote the total number of sequences for each of which $x_i=S_i\sigma_i^{-1}$ has been calculated. On H_0 each x_i approximates to a unit normal deviate. Hence, on H_0, the statistic

$$z=\sum_{i=1}^{N} x_i/\sqrt{N} \tag{4}$$

also approximates to a unit normal deviate. We propose z as the test statistic for the null hypothesis that the N protocols, considered as a group, show no overall trend in success probabilities. One can use either a one- or two-tailed significance test, depending on whether a directional trend has been forecast, since the test is equally sensitive to either an upward or downward trend in the observed response probabilities.

Table 1 gives a numerical example of the calculation of z for ten hypothetical, success–error sequences which reveal a small but consistent positive trend. The overall z is 2·38 which allows rejection of H_0 at the $P=0{\cdot}017$ level (two-tail). For comparison, the sequences were Vincentized into quartiles, with any odd trials being discarded from the middle of the sequence, and the Suppes–Ginsberg chi-square computed. This was 6·19 (3 df), giving $P > 0{\cdot}10$ on H_0. In this instance, z detects a small trend whereas the chi-square fails to do so.

TABLE 1. THE CALCULATION OF z FOR 10 SUCCESS–ERROR SEQUENCES

Protocol i	n_{1i}	n_{2i}	S_i	σ_i	σ_i^{-1}	x_i
1	2	3	4	3·46	0·289	1·155
2	1	5	1	3·16	0·316	0·316
3	4	2	4	3·74	0·267	1·070
4	7	5	3	12·30	0·081	0·244
5	2	4	2	4·31	0·232	0·464
6	3	3	1	4·58	0·218	0·218
7	3	4	4	5·65	0·177	0·708
8	2	2	4	2·58	0·388	1·550
9	5	3	9	6·70	0·149	1·343
10	1	3	1	2·24	0·446	0·446

$$\sum_{i=1}^{10} x_i = 7{\cdot}514, \qquad \sum_{i=1}^{10} \sigma_i^{-1} = 2{\cdot}563, \qquad N = 10.$$

$$z = \frac{\sum_{i=1}^{10} x_i}{\sqrt{N}} = \frac{7{\cdot}514}{\sqrt{10}} = 2{\cdot}38.$$

With correction for continuity,

$$z_c = \frac{1}{\sqrt{N}}\left[\sum_{i=1}^{10} x_i - \frac{\sum_{i=1}^{10} \sigma_i^{-1}}{N}\right] = 2{\cdot}30.$$

4. APPLICATIONS

Before entering upon a discussion of technical details, we wish to illustrate a few of the uses of the proposed z test for trend. The main restriction on usage is that the response categories can be ordered *a priori*, or that the responses be lumped into larger categories that are orderable. This restriction is necessary since S is otherwise indeterminate (i.e., each rearrangement of the rows of a $c \times n_i$ contingency table gives a different S value).

Success–error coding of sequences is an ordering, obtained by lumping all types of errors into a single category. Possibly a theory (say, of response learning in paired-associates) would lead to a further ordering within the errors themselves, e.g., ommissions, extralist intrusions, intralist intrusions, fragments of the correct response, or alternatively, the number of correct letters in a compound or trigram response. Or with well-learned responses, a theory may predict the trial ordering or ranked relative frequency over trials of particular types of error responses (e.g., Restle's (1964) model predicts that later errors are likely to be from particular confused items, whereas earlier errors are likely to be random guesses). These analyses can be done *conditionally*, deleting the successes and examining only the sequence of error categories for trends. The same analyses as before apply for any sub-sequence of the original sequence, since only the order of trials is important, though the analyses will not necessarily be independent.

Another example is confidence ratings (e.g., regarding the correctness of the response given). These can be treated in several ways: (*a*) the confidence ratings alone could be examined without reference to the correctness of the response given, (*b*) the sub-sequences of confidence ratings for error and correct response trials could be extracted and analysed separately, or (*c*) the confidence ratings could be ' unfolded ' in conjunction with the correctness of the response. For example, if the subject is allowed 5 confidence categories (e.g., $1=$lowest, $5=$highest confidence), then a correct response with confidence j is assigned score $+j$, whereas an error with that confidence is scored as $-j$. The scale from $+5$ to -5 is then ordered in something like ' relative strength of the correct response ' and the trend test can be applied.

Another example is a ' second-guess ' experiment (e.g. Binford and Gettys, 1965) where the three scores could be 1 if the first guess is correct, 2 if a wrong first guess is followed by a correct second guess, and 3 if both guesses are wrong. A further illustration is from an experiment by Bower (1967), where on each trial, to each paired-associate stimulus, the subject ranked all c available responses in terms of their likelihood of being the correct response to the stimulus. The score assigned for the trend test would be the ordinal position of the correct response in the trial-by-trial rankings given to a stimulus.

Metric data collected during learning provide further material for the trend test. This includes response latencies and response amplitudes, e.g., GSR, eyeblink, incidence of alpha rhythm in the EEG, etc. When these are recorded collateral to a primary response coded as success–error, or as CR–nonCR, then separate analyses of sub-sequences can be made, e.g., trends in latencies over error trials and over success trials, or trends in latencies on trials following error trials and following success trials. For such metric or continuous response measures, the number of categories is in practice simply those distinct values of the measure which occur in a particular protocol; in cases where all the scores are different, then all $n_{ij}=1$ and eqn. (1) for σ_i^2 can be simplified. S is calculated as before, as the sum over trials of the frequency of succeeding scores which exceed the given trial-score minus the frequency of succeeding scores which it exceeds.

Since S and z are relatively fast and easy to calculate in all the above cases, these statistics are recommended for their convenience, range of applicability, and power as tests for trend in such data. We now turn to some more detailed statistical considerations regarding the z test.

5. STATISTICAL PROPERTIES OF THE TEST

The following sections discuss the convergence of z to normality, a continuity correction, the power of z in detecting different types of trend, and its power efficiency with respect to the chi-square test. Specific calculations below are made for binary (success–error) sequences, but the general results also hold when the test is applied to multiple response categories. Moreover, though our investigations of convergence and power are carried out only for 'short' sequences of three to six trials, the results also apply to longer sequences, since the approximation of S to normality increases with sequence length.

Approximation of z to normality. Though the proposed test utilizes the approximation of z to normality, this can be avoided, in principle, by explicitly convoluting the N individual exact distributions of S_i (or x_i) and thus obtaining the exact distribution of z. However, the labour involved makes this exact procedure impractical and is an argument for the convenient normal approximation.

The goodness of this approximation of z to the unit normal distribution depends in general both on the lengths of the protocols (the n_i's), and on the number of protocols (N), which are pooled. The speed of convergence to normality may be investigated by comparing the first few cumulants of the distribution of z to those of the unit normal.

For convenience in the following, let κ_{ri} denote the rth cumulant of x_i, γ_r denote the rth cumulant of z, and $m_i = n_{i_1} n_{i_2}$. On the null hypothesis the cumulants of x_i for a dichotomy are as follows (cf. the formulae for the general case of S in Jonckheere, 1954 *a*, *b*); all the odd-numbered cumulants are zero, since the distribution of S is symmetric, while

$$\kappa_{2i} = 1,$$

$$\kappa_{4i} = -\frac{6}{5}\left[\frac{n_i}{m_i} - \frac{1}{(n_i+1)}\right],$$

and

$$\kappa_{6i} = \frac{48}{7(n_i+1)^2} + \left[\frac{24n_i}{7m_i(n_i+1)}\right]\cdot\left[\frac{(2n_i^2+2n_i-1)}{m_i} - \frac{(4n_i+5)}{(n_i+1)}\right]. \qquad (5)$$

Hence, the odd-numbered cumulants of z are zero, while

$$\gamma_2 = 1,$$

$$\gamma_4 = \frac{1}{N^2}\sum \kappa_{4i},$$

and

$$\gamma_6 = \frac{1}{N^3}\sum \kappa_{6i}.$$

Compared to the unit normal, z has the same first three moments, γ_4 converges to zero at the rate of N^{-1}, γ_5 is zero, while γ_6 converges to zero at the rate of N^{-2}. Also if $n_i \geq 3$ for all i, max $\gamma_4 = -1{\cdot}5/N$ and max $\gamma_6 = 9{\cdot}75/N^2$. Thus, the approximation of z to the unit normal distribution is quite close, and converges rapidly as N, the number of protocols, is increased.

Continuity correction S_i is a discrete random variable with successive values occurring in steps of two. If $m = n_1 n_2$ is even, S has the possible values $\pm 2i$ for $i = 0, 1, \ldots, m/2$. If m is odd, S has values $\pm(1+2i)$ for $i = 0, 1, \ldots, (m-1)/2$.

Therefore, when the m_i are fixed, the sum of N of the S_i has possible values which jump by two's. When the exact distribution of $\sum S_i$ is approximated by the continuous normal, the discreteness of $\sum S_i$ slightly affects the tail of the distribution which is important for assessing the level of significance of particular results. In the simplest case, when all m_i and n_i (and hence σ_i) are equal, z is a constant times $\sum S_i$, and a correction for continuity is effected in this case by reducing the numerical value of $\sum S_i$ by 1 before computing z. When the n_i 's (and hence σ_i's) are different for the various protocols, an appropriate correction for continuity would seem to be

$$z_c = \frac{1}{\sqrt{N}}\left\{\left|\sum \frac{S_i}{\sigma_i}\right| - \frac{1}{N}\sum \frac{1}{\sigma_i}\right\}. \tag{6}$$

If all σ_i were equal to a constant σ, then eqn. (6) reduces to

$$z_c = \frac{|\sum S_i| - 1}{\sigma\sqrt{N}}.$$

Of course, as N becomes large the correction factor becomes negligible and can be disregarded.

Figure 1 illustrates the value of the continuity correction when N is small (4 in this instance). The distribution of S for two successes and two errors was convoluted four times to obtain the exact probability distribution of $\sum S_i$

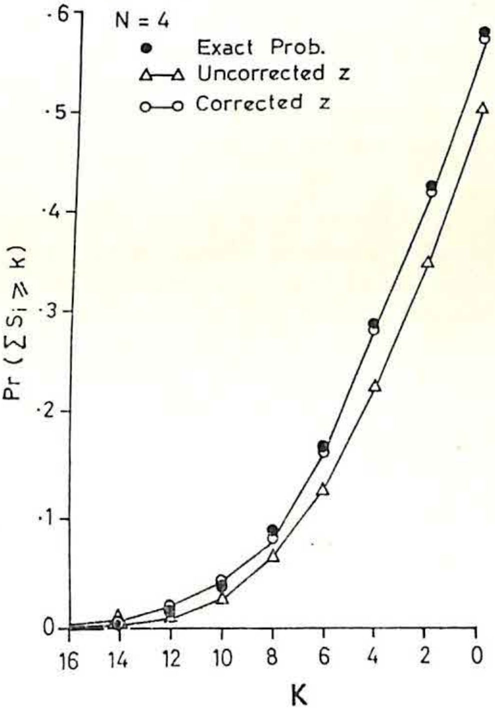


FIGURE 1. Approximation of corrected and uncorrected z to exact tail probabilities.

(or z, since σ_i is constant here). The tail probabilities, $P(z > K)$, for the uncorrected and corrected z were compared to those of the normal deviate. Figure 1 shows that the exact tail probabilities are well approximated by the unit normal function when its argument is the corrected z, but not when its argument is the uncorrected z. Thus, when N is small, the continuity correction materially aids the normal approximation to the exact probabilities.

6. POWER OF THE z-TEST

Support for the null-hypothesis of stationarity can always be obtained by using a small enough sample of short sequences. Only by assessing the power of the statistical tests employed can we avoid this facile acceptance of the null-hypothesis, and hence spurious support for the learning model under consideration.

For our case, H_0 is the null-hypothesis of no trend; in particular given the marginal totals n_{i1} and n_{i2}, H_0 states that all distinct permutations of n_{i1} successes and n_{i2} errors are equally likely. The general alternative is that there is, say, an upward trend; in particular, that sequences with high S-values (with successes clustered at the end) have relatively high probabilities.

As often happens with non-parametric tests, the basic problem in calculating their power is the difficulty of deriving the distribution of the test statistic (z here) given only general specification of the non-null hypothesis, when the equally likely restriction is removed. When equal probability of all distinct permutations of categories of response is denied, then the non-null distribution of z has to be worked out for each class of alternative hypotheses.

We now investigate the power of z for two specific trend schemes, one which generates a linear trend and one which generates a logistic trend.

A linear trend scheme The power of the z-test will first be evaluated against a one parameter scheme which generates a linear upward trend with equal trial-to-trial increments in the probability of success. Such a scheme will always provide a fair approximation to any monotonic trend observed in practice.

Given n_1 successes in n trials, the null hypothesis, H_0, assigns a probability of $\binom{n}{n_1}^{-1}$ to each possible sequence. The alternative hypothesis, H_d, considered here is that the probability of a given sequence increases linearly with its S value, viz.,

$$Pr\{\text{sequence with value } S\} = \binom{n}{n_1}^{-1} + \frac{d}{2}S. \tag{7}$$

The trend parameter is d; H_0 is that $d=0$; whereas H_d is that $d>0$. Eqn. (7) constrains the value of d since the probability must lie between 0 and 1. Since the maximum and minimum values of S are $\pm n_1 n_2$, the constraints on d can be derived from eqn. (7).

Consider a simple example, namely, one success and two errors. There are three sequences possible, corresponding to the success occurring on trial 1, 2 or 3. The S values assigned to these three sequences are -2, 0, $+2$, respectively, and their probabilities on H_d are $(1/3)-d$, $(1/3)$, and $(1/3)+d$ respectively (where $d \leqslant 1/3$). In this simple case, the trend can be read off directly from the sequence probabilities; in general the proportion of successes on trial i is obtained by summing the probabilities of those sequences having a success on trial i.

Several simple cases are summarized in Table 2. The left-hand column, labelled n_1/n_2, gives the number of successes and errors. The next columns give the trend in proportions of successes over trials, while the last column gives the constraint on d implied by eqn. (7).

TABLE 2. TREND FOR VARIOUS CASES OF n_1 SUCCESSES AND n_2 ERRORS

	TRIALS						
n_1/n_2	1	2	3	4	5	6	$d \leqslant x$
1/1	$0{\cdot}5-0{\cdot}5d$	$0{\cdot}5+0{\cdot}5d$					1·000
1/2	$0{\cdot}33-d$	0·33	$0{\cdot}33+d$				0·333
2/2	$0{\cdot}5-3d$	$0{\cdot}5-d$	$0{\cdot}5+d$	$0{\cdot}5+3d$			0·083
2/3	$0{\cdot}4-6d$	$0{\cdot}4-3d$	0·4	$0{\cdot}4+3d$	$0{\cdot}4+6d$		0·033
3/3	$0{\cdot}5-15d$	$0{\cdot}5-9d$	$0{\cdot}5-3d$	$0{\cdot}5+3d$	$0{\cdot}5+9d$	$0{\cdot}5+15d$	0·011

The cumulants of the distribution of S under H_d may be obtained as follows. From eqn. (7) we have

$$Pr\{S_i=S|H_d\}=Pr\{S_i=S|H_0\}\ (1+\alpha S),$$

where $\alpha=\binom{n}{n_1}\frac{d}{2}$. Hence, if $G_d(x)$ and $G_0(x)$ are the probability generating functions of S on H_d and H_0 respectively,

$$G_d(x)=G_0(x)+\alpha x G_0'(x).$$

Putting $x=e^{it}$, we obtain a corresponding equation relating the characteristic functions of S on H_d and H_0;

$$\phi_d(t)=\phi_0(t)+\alpha\phi_0'(t).$$

Differentiating and putting $t=0$, we can now express the raw moments, $\mu'_i(d)$, of S on H_d in terms of the raw moments $\mu_i'(0)$, of S on H_0. Thus,

$$\mu_r'(d)=\mu_r'(0)+\alpha\,\mu'_{r+1}(0) \qquad (r=0,\ 1,\ \ldots).$$

In addition, since the distribution of S on H_0 is symmetrical we have

$$\mu'_{2r+1}(0)=0.$$

Hence, if $K_i(d)$ and $K_i(0)$ are the ith cumulants of the distributions of S on H_d and H_0 respectively, we have

$$\begin{aligned}
&K_1(d)=\alpha K_2(0),\\
&K_2(d)=K_2(0)\ \{1-\alpha^2K_2(0)\},\\
&K_3(d)=\alpha\{K_4(0)+2\alpha^2K_2{}^3(0)\},\\
&K_4(d)=K_4(0)-2\alpha^2K_2(0)\{2K_4(0)+3\alpha^2K_2{}^3(0)\}.
\end{aligned}$$

Assuming all N protocols to have n_1 successes and n_2 errors, then the rth cumulant of z on H_d will be

$$\lambda_r = N^{\{(2-r)/2\}} \kappa_2^{-r/2}(0)\ \kappa_r(d),$$

where

$$\kappa_2(0) = (n_1 n_2/3)\ (n+1).$$

And in particular, we have for the mean and variance:

$$\lambda_1 = \alpha\sqrt{\{N\kappa_2(0)\}} = Ad\sqrt{(N)},$$

$$\lambda_2 = \sqrt{\{1-\alpha^2\kappa_2(0)\}} = \sqrt{(1-A^2 d)},$$

where

$$A = \binom{n}{n_1}\sqrt{\left\{n_1 n_2 \frac{(n+1)}{12}\right\}}.$$

The approximation to the cumulants of a normal distribution appears to be adequate, even for the simple example given above where $n_1 = 1$ and $n_2 = 2$. Putting $d = 0{\cdot}10$, giving a three-trial trend of probabilities 0·23, 0·33, 0·43, then $\lambda_1 = 0{\cdot}245\sqrt{N}$, $\lambda_2 = 0{\cdot}94$, $\lambda_3 = -0{\cdot}338/\sqrt{N}$, and $\lambda_4 = -1{\cdot}16/N$. Even for $N = 10$ the third and fourth cumulants of z are close enough to zero to warrant using a normal approximation for its distribution.

Hence the power of the z-test against the alternative H_d is given by

$$Pr\{z \geqslant C_\alpha | H_d\} = G(U),$$

where

$$U = \frac{C_\alpha - \lambda_1}{\sqrt{\lambda_2}} = \frac{C_\alpha - Ad\sqrt{(N)}}{\sqrt{(1-A^2 d)}}.$$

C_α is the unit-normal deviate giving an α-sized critical region (e.g., $C_{0{\cdot}05} = 1{\cdot}645$ for a one-tailed test), and $G(U)$ is the probability that the unit normal deviate is greater than or equal to U.

It can be seen that the power of the test increases with the number, N, and the length, n, of the protocols, and also with the size of the linear trend parameter, d.

Power of the z-test against logistic alternatives Since several response-strength theories of learning (Luce, 1959; Spence, 1956) imply logistic learning curves, we now discuss this type of trend as an alternative to the null hypothesis. Statistics suitable for the detection of such trends have been investigated by Cox (1958), and Barton, David and Mallows (1958), and their relevant results may be briefly presented.

Let t_i equal 1 or 0 according to whether a subject has a success or an error on trial i. The logistic trend for the increase in probability of success is given by the following expression:

$$Pr\{t_i = 1\} = \frac{A\theta^i}{1 + A\theta^i},$$

where A determines the initial probability of success, while θ determines the increase in the probability of success as a function of the trial number. When $\theta = 1$, we have the null hypothesis that there is no trend in the probability of success as a function of trial number.

The likelihood of obtaining a particular sequence of n responses $t_1, \ldots, t_n$ is given by

$$L_n = \frac{A^{n_1}\theta^{T}}{\prod_{i=1}^{n} (1+A\theta^{i})},$$

where $n_1 = \sum_{i=1}^{n} t_i$, the total number of successes, and $T = \sum_{i=1}^{n} i\, t_1$, the summed ranks (trial numbers) of the successes. Since $T = \frac{S}{2} + \frac{n_1}{2}(n+1)$, where S is Kendall's measure used above, we now have

$$L_n = \frac{A^{n_1}\, \theta^{n_1(n+1)/2}\, \theta^{S/2}}{\prod_{i=1}^{n} (1+A\theta^{i})}.$$

Let $N_{S,\,n_1}$ be the number of permutations of n_1 successes and $n-n_1$ errors which give the same value of S. The probability of obtaining a particular value of S with n_1 successes out of n trials is

$$Pr\{S, n_1, n\} = L_n N_{S,\,n_1}. \tag{8}$$

Hence the probability of n_1 successes occurring is given by

$$Pr\{n_1\} = \sum_{S} L_n N_{S,\,n_1},$$

where the summation is over all possible values of S. Thus the probability of obtaining a value of S conditional upon n_1 successes occurring in n trials is, after reduction,

$$Pr\{S|n_1, n, \theta\} = \frac{N_{S,\,n_1}\theta^{S/2}}{\sum_{S} N_{S,\,n_1}\theta^{S/2}}. \tag{9}$$

Using the known generating function for $N_{S,\,n_1}$, it can be shown that

$$\sum_{S} N_{S,n_1}\theta^{S/2} = \theta^{-n_1 n_2/2} \prod_{i=1}^{n_2} \left(\frac{1-\theta^{n_1+i}}{1-\theta^{i}}\right),$$

and hence that

$$Pr\{S|n_1, n, \theta\} = N_{S,\,n_1}\theta^{(n_1 n_2+S)/2} \prod_{i=1}^{n_2} \left(\frac{1-\theta^{i}}{1-\theta^{n_1+i}}\right). \tag{10}$$

When $\theta = 1$, we have the probability distribution of S under H_0, viz.,

$$Pr\{S|n_1, n, \theta=1\} = N_{S,\,n_1} \Big/ \binom{n}{n_1}. \tag{11}$$

The probability distribution of S under H_θ is easily obtained from eqn. (11): one simply multiplies the probability of S under H_0 by $\theta^{S/2}$, and divides by a factor that makes the probabilities sum to one. In comparison with the previous scheme, whereas the linear trend added $dS/2$ to each probability of S arising under H_0, the logistic scheme multiplies each probability by amounts proportional to $\theta^{S/2}$.

We can also easily obtain the cumulant generating function of the distribution of S under H_θ as follows. Let $G(x;\theta)=\sum_S Pr\{S|n_1, n, \theta\}x^S$ be the probability generating function of S under H_θ. From eqns. (9) and (11) we have

$$Pr\{S|n_1, n, \theta\}=\frac{\binom{n}{n_1} Pr\{S|n_1, n, \theta=1\}\,\theta^{S/2}}{\sum_S \binom{n}{n_1} Pr\{S|n_1, n, \theta=1\}\theta^{S/2}},$$

and thus

$$G(x;\theta)=\frac{G(\theta^{1/2}x;\theta=1)}{G(\theta^{1/2};\theta=1)}. \tag{12}$$

Putting $x=e^t$, $\theta=e^{2\alpha}$ and taking logarithms we obtain

$$\psi(t;\theta)+\log G(e^\alpha;\theta=1)=\psi(t+\alpha;\theta=1), \tag{13}$$

an expression relating the cumulant generating functions $\psi(t;\theta)$ and $\psi(t+\alpha;\theta=1)$, of the distribution of S under H_θ and H_0. Thus, if $\kappa_i(\theta)$ and $\kappa_i(1)$ stand for the ith cumulants of S under H_θ and H_0, we have

$$\sum_{i=1}^{\infty} \kappa_i(\theta)\frac{t^i}{i!}=\sum_{i=1}^{\infty}\frac{\kappa_i(1)}{i!}(t+\alpha)^i .$$

Equating the values of t, the cumulants of S under H_θ may be expressed in terms of those under H_0. In general we have

$$\kappa_i(\theta)=\sum_{r=i}^{\infty}\frac{\alpha^{r-i}}{(r-i)!}\kappa_r(1),$$

and recalling that the odd cumulants of S under H_0 are zero, we have in particular that

$$\kappa_{2i+1}(\theta)=\alpha\kappa_{2i+2}(1)+\frac{\alpha^3}{3!}\kappa_{2i+4}(1)+\frac{\alpha^5}{5!}\kappa_{2i+6}(1)+\ldots$$

$$\kappa_{2i}(\theta)=\kappa_{2i}(1)+\frac{\alpha^2}{2!}\kappa_{2i+2}(1)+\frac{\alpha^4}{4!}\kappa_{2i+4}(1)+\ldots$$

where $i=0, 1, \ldots$ and $\alpha=\log\theta/2$.

Thus in so far as the distribution of S on the null hypothesis, H_0, approaches normality quite rapidly, the distribution of S on the alternative hypothesis H_θ also rapidly converges to a normal distribution for small values of θ. Hence the distribution of z on H_θ will be well approximated by a normal distribution, even for small values of n_i (the lengths of the individual protocols), provided N (the number of protocols) is reasonably large.

We may also easily obtain explicit expressions for the mean and variance of S under H_θ, since the probability generating function of S under H_0 is

$$G(x;\,\theta=1)=\prod_{i=1}^{n_2}\left[\frac{x^{n_1+i}-x^{-(n_1+i)}}{x^i-x^{-i}}\right]\Big/\binom{n}{n_1}.$$

Hence putting $x=e^t$ and $\theta=e^{2\alpha}$, taking logarithms, and substituting in eqn. (12) we obtain an explicit form for the cumulant generating function $\psi(t;\,\theta)$. Differentiating and putting $t=0$, we have in particular

$$\kappa_1(\theta)=\sum_{i=1}^{n_2}\Big[(n_1+i)\coth\,\{\alpha(n_1+i)\}-i\coth\,(\alpha i)\Big],$$

$$\kappa_2(\theta)=\sum_{i=1}^{n_2}\left[\frac{i^2}{\sinh^2(\alpha i)}-\frac{(n_1+i)^2}{\sinh^2\{\alpha(n_1+i)\}}\right],$$

where $\alpha=\log\theta/2$;

or

$$\kappa_1(\theta)=\sum_{i=1}^{n_2}\left[(n_1+i)\left(\frac{\theta^{n_1+i}+1}{\theta^{n_1+i}-1}\right)-i\left(\frac{\theta^i+1}{\theta^i-1}\right)\right],$$

and

$$\kappa_2(\theta)=4\sum_{i=1}^{n_2}\theta^i\left[\left(\frac{i}{\theta^i-1}\right)^2-\theta^{n_1}\left(\frac{n_1+i}{\theta^{n_1+i}-1}\right)^2\right]. \tag{14}$$

To illustrate the adequacy of a normal approximation to the distribution of S for a special case of H_θ, consider sequences with two successes and two errors. The non-null distribution of S, obtained from eqn. (12), is shown in Table 3.

TABLE 3. DISTRIBUTION OF S FOR TWO SUCCESSES AND TWO ERRORS UNDER H_θ [FROM EQN. (13)]

S	N_S	$\theta^{(s+4)/2}$	$Pr\{S\}$
4	1	θ^4	$A\theta^4$
2	1	θ^3	$A\theta^3$
0	2	θ^2	$2A\theta^2$
-2	1	θ	$A\theta$
-4	1	1	A

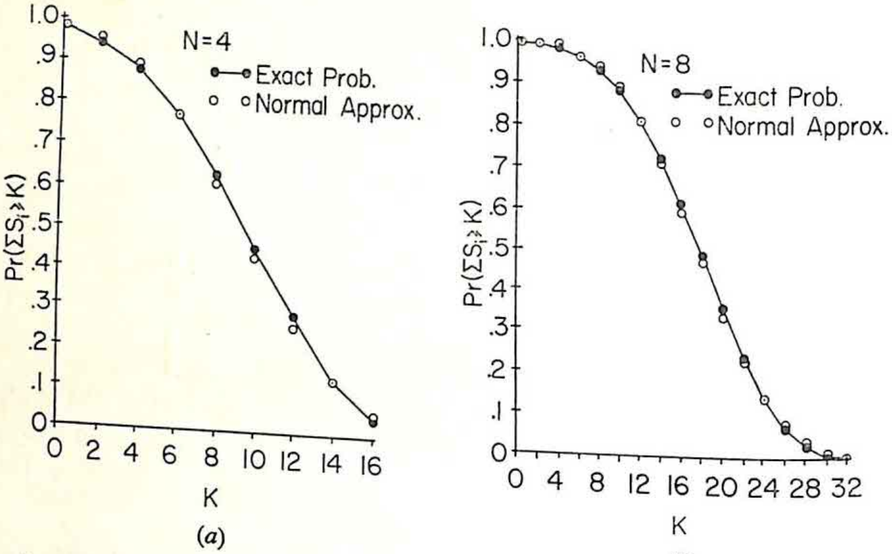
where $A=(1-\theta)\,(1-\theta^2)/(1-\theta^3)\,(1-\theta^4)$.

The mean and variance are given by eqns. (14), with $n_1=n_2=2$. If $\theta=2$, for example,

$$\kappa_1(2)=\frac{72}{35}=2{\cdot}06,\quad\text{and}\quad\kappa_2(2)=\frac{5736}{1225}=4{\cdot}68\,.$$

The distribution of S for $\theta=2$ was convoluted $N=4$ and $N=8$ times to determine the exact probabilities of the distribution of z. For fixed n_1 and n_2, z approaches normality if ΣS_i does. Hence, the exact probability that ΣS_i exceeded K was compared to that calculated (with continuity correction) on the assumption that ΣS_i was normally distributed with mean $N\kappa_1(2)$ and variance

$N\kappa_2(2)$. The comparisons of tail probabilities for $\theta=2$ and for $N=4$ and 8 are graphed in Fig. 2. The normal approximation is already quite close for $N=4$ (maximal discrepancy of 0·025 at $K=10$) and is much improved for $N=8$. The $\theta=2$ case is useful to examine in this respect since, within the reasonable range of θ values, this one induces the maximal skewing of the S distribution. The normal approximation for smaller $|\theta-1|$ values, or for larger n_1 or N, may therefore be expected to be even better than in Figure 2.



(*a*) (*b*)

FIGURE 2. Adequacy of z approximation for special cases: (*a*) $\theta=2, N=4$; (*b*) $\theta=2, N=8$.

Hence if all N protocols have n_1 successes and n_2 errors, the distribution of z on the hypothesis of a logistic trend can safely be assumed to be normally distributed, the mean and variance being

$$\lambda_1=\frac{\kappa_1(\theta)}{\sigma}\sqrt{N}$$

and

$$\lambda_2=\frac{\kappa_2(\theta)}{\sigma},$$

where $\kappa_1(\theta)$ and $\kappa_2(\theta)$ are given by eqn. (14) and $\sigma^2=n_1n_2(n+1)/3$. Thus the power of the z test against this type of trend will be $G(U)$, where

$$U=\frac{C_\alpha\sigma-\sqrt{(N)}\kappa_1(\theta)}{\sqrt{\kappa_2(\theta)}}.$$

C_α is the unit normal deviate giving an α-sized critical region, and $G(U)$ is the probability that the unit normal deviate is greater than or equal to U.

It should be noted that for both the linear and the logistic trend schemes discussed above, the probability of success at any trial is considered independent of the outcome on previous trials; and hence these alternatives do not represent cases where probability changes are contingent upon the prior sequence of responses made by the subject.

Relative efficiency of z and chi-square In this section we attempt an approximate assessment of how much more efficient is the z-test for trend in comparison to the chi-square tests in current use. The relative power efficiency of the z and chi-square tests may be roughly estimated as follows. Selecting values of n_1, n_2, and the trend parameter (d or θ), let N_z and N_Q be the number of protocols required to have power P when using the z and chi-square tests for trend respectively. Then the ratio N_z/N_Q estimates the efficiency of z relative to the chi-square for this specific case. This ratio generally can be expected to be a function of the n_i, the trend parameter, and the power-level of the comparison. In the following, we shall use the 50 per cent power level for comparisons at a significance level of $\alpha = 0{\cdot}05$. Calculations of power for the chi-square statistic are complicated by the fact that under the non-null hypothesis the usual statistic is distributed as a *non-central* chi-square, with a non-centrality parameter depending upon N and the trend parameter. Strict general solutions are difficult to obtain and we shall employ the following approximate methods. Suppose a chi-square with $k-1$ df has an $\alpha = 0{\cdot}05$ significance value of R. By setting observed frequencies in the contingency table equal to those theoretically expected on the trend hypothesis and then calculating the chi-square, we obtain a rough estimate (which improves as N becomes larger) of the mean of the non-central chi-square. This estimated mean will be a function of the number of protocols, N, and the trend parameter, d or θ. Because of the skew of chi-square distributions (i.e., the median is less than the mean), if we set this estimated mean equal to R and solve for N, then the power for this value of N is somewhat less than $0{\cdot}50$ (i.e., the probability area above R is less than $0{\cdot}50$). Thus, by putting this value of N equal to N_Q we shall be giving the chi-square a slight advantage as compared to the z test when estimating the 50 per cent power efficiency.

A final word before beginning the comparison. The trend schemes illustrated assign probabilities to entire sequences of successes and errors, and they do not re-assign that probability independently to each response in the sequence. Hence, with sequences k trials long, one has N protocols each with k *dependent* observations, not Nk independent observations. Thus, Cochran's modification of the chi-square for k related observations is the appropriate test statistic for our trend schemes (cf. Cochran, 1950; Siegel, 1956). In the Suppes and Ginsberg (1963) paper, Cochran's test was not used because the particular null model being tested assumed that successive responses were independent. However, it would seem preferable not to have decisions about trends dependent upon assuming or demonstrating independence of successive responses. Thus, Cochran's test would seem generally to have wider applicability, as well as being required by our specific comparisons.

Cochran's Q statistic is defined for N subjects (sequences) each yielding k related observations (trials) as follows:

$$Q_{k-1} = \frac{k(k-1)\sum_{j=1}^{k}(G_j - G)^2}{k\sum_{i=1}^{N} L_i - \sum_{i=1}^{N} L_i^2},$$

where L_i is the number of successes for the ith protocol (our n_{i1}), k is the number of trials, G_j is the total number of successes on trial j, and G is the mean number of successes per trial. McNemar's test (cf. Siegel, 1956) is a special case of Q when $k=2$. On the null hypothesis of equal proportions of successes on the k trials, Q is approximately distributed as chi-square with $k-1$ degrees of freedom.

For our calculations, we fix the number of successes at n_1 and trials at $k=n$, replace L_i by n_1, and set G_j equal to its expected value of Np_j, where p_j is the mean success probability on trial j (cf. Table 2). We thus obtain a rough estimate of the mean of Q, which will be a function of N and the trend parameter. Equating this to R and solving yields an estimate of N_Q.

We will illustrate this procedure for the linear trend for the 1/2 case from Table 2, with $p_1 = 0{\cdot}33 - d$, $p_2 = 0{\cdot}33$, and $p_3 = 0{\cdot}33 + d$. With appropriate substitutions, Q in this case reduces to

$$Q_2 = \frac{2n(n-1)Nd^2}{n_1(n-n_1)} = 6Nd^2.$$

With 2 df, a chi-square value above 5·99 is significant at the 5 per cent level. Hence, for given d, the N required to make $Q_2 \geqslant 5{\cdot}99$ is

$$N_Q \geqslant \frac{5{\cdot}99}{6d^2} \simeq \frac{1}{d^2}.$$

Proceeding analogously with the z-test for this case we have, for $\alpha = 0{\cdot}05$ with a two-tailed test,

$$2{\cdot}46d\sqrt{N} \geqslant 1{\cdot}96,$$

$$N_z \geqslant 0{\cdot}634/d^2.$$

For this case, as for all others examined with the linear-trend scheme, N_Q and N_z turn out to be just a different constant divided by d^2. Hence, the efficiency of Q relative to z is the ratio, N_z/N_Q, which gives the proportion of sample size required when z rather than Q is used to detect a trend. This index is given in Table 4 for the simpler cases of n_1 successes and n_2 errors. This ratio is independent of d, but this follows from our choice of the 50 per cent power level for comparison of the two tests.

As Table 4 shows, for $n_1 = n_2 = 1$, z and Q are equivalent tests with the same power; however, the efficiency of Q relative to z decreases as the sequences become longer. This is to be expected, since S then has a larger range over which to vary. The indices in Table 4 are slight over-estimates of the true relative efficiency since it will be recalled that the chi-square power is actually somewhat less than 0·50. If in addition a directional trend is predicted, enabling a one-tailed test with z, then the relative efficiency of z improves even further. The ratios in Table 4 may then be reduced by multiplication by $0{\cdot}705 = (C_{0\cdot10}/C_{0\cdot05})^2$. The chi-square, of course, is inherently a two-tailed test.

TABLE 4. EFFICIENCY OF COCHRAN'S Q RELATIVE TO z IN TESTING FOR A LINEAR TREND WITH n_1 SUCCESSES AND n_2 ERRORS

n_1/n_2	N_z/N_Q
1/1	1·000
1/2	0·634
2/2	0·490
2/3	0·405
3/3	0·347

TABLE 5. NUMBER OF 2/2 CASES REQUIRED BY z OR Q TEST TO HAVE PROBABILITY 0·50 OF DETECTING A LOGISTIC TREND WITH PARAMETER θ

θ	*Average increase per trial*	N_z	N_Q
1·1	0·025	254·0	483·0
1·2	0·045	70·4	144·7
1·3	0·063	35·0	72·0
1·4	0·081	21·4	44·3
1·5	0·096	15·2	31·3
1·6	0·110	11·5	24·7

A somewhat different approach is illustrated for the logistic trend in Table 5. It shows the N_z and N_Q required to have 50 per cent power when there are two successes and two errors with sequences generated by the logistic scheme with parameter θ (cf. Table 3). These use an $\alpha = 0{\cdot}05$, two-tail rejection region. The four trial-proportions, p_j, have a mean of 0·50; the average trial-to-trial increment as a function of θ is indicated in the second column of Table 5. The efficiency of Q relative to z in this instance is approximately 0·50, independent of θ. This is reminiscent of the results in Table 4 with the linear trend, where relative efficiency was independent of d. Interestingly, Table 4 also gave a relative efficiency near 0·50 (0·49 actually) for the 2/2 case and a linear trend. More cases will have to be examined, but on these preliminary grounds it looks as though relative efficiency of z to Q at 50 per cent power is primarily a function of protocol length and is unrelated to amount of trend (d or θ) and type of trend (linear or logistic).

7. HOMOGENEITY OF SUBJECTS' TRENDS

It is perhaps obvious that without a test of inter-subject homogeneity, non-significance of any overall test of trend for a sample of protocols does not necessarily entail consistency with a learning model predicting stationarity of individual sequences. The z test is for overall group trend and, like the chi-square, could fail to reach significance if different protocols show strong trends, but in opposite directions.

The following test for the homogeneity of trends is a useful safeguard when interpreting the outcome of the z test. Its only drawback is that the statistic employed converges rather slowly to a chi-square distribution, and thus its use is restricted to longer protocol sequences than those which can be used with the z test.

It is reasonable that a test for subject homogeneity should depend upon the dispersion of the individual x_i about the observed mean $\bar{x}$. This suggests developing a test along the following lines. Since under H_0 each x_i approaches to a unit normal deviate in the limit as n_{i1} and n_{i2} become large, it follows that x_i^2 will approach to a chi-square distribution with one degree of freedom under the same limiting conditions. On independence assumptions, the sum of the x_i^2 over the N independent protocols will thus approach to a chi-square distribution with N degree of freedom. Thus $\sum x_i^2$ is not an appropriate test statistic for individual differences, however, since it can be large simply because the mean, $\bar{x}$, is large. But we can subtract a correction factor for the mean. Since z approaches the unit normal distribution, z^2 approaches to a chi-square distribution with one degree of freedom. It therefore follows that

$$W=\sum x_i^2-z^2=\sum x_i^2-N\bar{x}^2=\sum (x_i-\bar{x})^2 \tag{15}$$

should approximate to the chi-square distribution with $N-1$ df as each n_{i1} and n_{i2} becomes large. This W-statistic, which is just the sum of squared deviations of the x_i about their mean, is proposed as an appropriate test statistic for the null hypothesis of homogeneity of subjects' trends: it is independent of the z test for group trend, in so far as the x_i are normally distributed, and so any pattern of significances is possible among the two statistics W and z.

The only difficulty with this test is that under H_0, x_i^2 converges very slowly to a chi-square distribution as n_{i1} and n_{i2} increase, and each n_{ij} must be 'fairly large' (say 10 or more) before the chi-square approximation is very good. This poor convergence of x_i^2 is not ameliorated by summing many of them into the W statistic, since the sum of many non-chi-square variates does not converge to chi-square. After showing these convergence properties, we turn to the question of the approximation of W (or a modification of W) to familiar distributions.

Convergence properties Convergence will be examined in terms of the theoretical cumulants of x_i^2 and W under H_0. A chi-square variate with $N-1$ degrees of freedom has its first cumulant equal to $N-1$, its second to $2(N-1)$, its third to $8(N-1)$, and its fourth to $48(N-1)$. Knowing the theoretical cumulants of x_i, tedious algebra assisted by Fisher's k-statistics (Kendall, 1958, p. 281) enables one to derive the cumulants for x_i^2 and hence for W. The first four cumulants of W will be denoted as π_1 to π_4, and prove to be:

$$\begin{aligned}
\pi_1 &= N-1,\\
\pi_2 &= 2(N-1)+\left(\frac{N-1}{N}\right)^2 \Sigma\kappa_{4i},\\
\pi_3 &= 8(N-1)+12\left(\frac{N-1}{N}\right)^2 \Sigma\,\kappa_{4i}+\left(\frac{N-1}{N}\right)^3 \Sigma\,\kappa_{6i},\\
\pi_4 &= 48(N-1)+144\left(\frac{N-1}{N}\right)^2 \Sigma\,\kappa_{4i}+24\left(\frac{N-1}{N}\right)^3 \Sigma\,\kappa_{6i}\\
&\quad+\left(\frac{N-1}{N}\right)^4 \Sigma\,\kappa_{8i}+\frac{8(N-1)}{N^4}(4N^2-9N+6)\left(\Sigma\,\kappa_{4i}\right)^2,
\end{aligned} \tag{16}$$

where κ_{4i} and κ_{6i} are as given in eqn. (5), and κ_{8i} is the eighth cumulant of x_i, not presented here (cf. Jonckheere, 1954) due to its length. Examination of these expressions shows that these cumulants approach those of a chi-square only if the extra summation terms on the right add to zero. But it is unlikely that these extra terms effectively sum to zero, except when *each* n_{i_1} and n_{i_2} becomes large. For example, consider π_2: as the number of protocols, N, becomes large, the extra term is essentially $\Sigma\,\kappa_{4i}$. But from eqn. (5) it is known that κ_{4i} is always negative, converging to zero only as n_{i_1} and n_{i_2} become large. Hence π_2 in practice will always be less than the second cumulant for a chi-square variate, and the more so, the shorter the sequences involved. Similar remarks apply to π_3 and π_4. The cumulants of the individual protocols enter in such a manner that summing N of them does not eliminate disturbances introduced by a few short sequences. This leads to an obvious caution regarding use of W as a test statistic: W should not be used with short sequences (say, n_{i_1} or n_{i_2} less than ten) since it is not then exactly distributed as a chi-square statistic.

The best approximation to the exact distribution of W for small values of n_i appears to be obtained by normalizing W, using a non-linear transformation derived from Edgeworth's Type A expansion of frequency functions (cf. Kendall, 1958, pp. 163–166).

If we put

$$\omega_1=\frac{W-(N-1)}{\sqrt{\pi_2}} \quad \text{and} \quad L_3=\pi_3/\pi_2^{3/2},$$

then

$$\xi=\omega_1-\frac{L_3}{6}(\omega_1^2-1)$$

will be approximately a unit normal deviate for even small values of n_{i_1} and n_{i_2}. Table 6 shows the accuracy of this approximation for selected values of N, n_{i_1}, n_{i_2}, and W, where the exact values of $Pr\{W\geqslant W_0\}$ were obtained by convoluting the distribution of W for $N=10$ and 16.

TABLE 6. ACCURACY OF NORMALIZED W COMPARED WITH EXACT PROBABILITIES

N	n_{i1}	n_{i2}	W_0	$Pr\{W \geqslant W_0\}$	ξ_0	$Pr\{\xi \geqslant \xi_0\}$
10	3	1	15·12	0·007	2·30	0·011
			14·48	0·024	2·06	0·020
			13·20	0·048	1·58	0·057
10	2	2	16·86	0·011	2·24	0·013
			15·66	0·025	1·94	0·026
			14·64	0·049	1·67	0·048
16	3	1	22·40	0·010	2·24	0·013
			21·60	0·025	2·00	0·023
			20·60	0·056	1·70	0·045
16	2	2	25·05	0·010	2·32	0·010
			23·36	0·026	1·96	0·025
			21·86	0·051	1·64	0·051

TABLE 7. VALUES OF κ_{4i} AND κ_{6i} [EQN. (5)] FOR $n_i = n_{i1} + n_{i2} \leqslant 10$ AND $m_i = n_{i1} n_{i2}$

n_i	m_i	κ_{4i}	κ_{6i}
2	1	−2·00	16·00
3	2	−1·50	9·75
4	3	−1·36	8·32
	4	−0·96	4·08
5	4	−1·30	7·75
	6	−0·80	2·89
6	5	−1·27	7·46
	8	−0·73	2·43
	9	−0·63	1·80
7	6	−1·25	7·29
	10	−0·69	2·20
	12	−0·55	1·39
8	7	−1·24	7·19
	12	−0·67	2·07
	15	−0·51	1·19
	16	−0·47	1·00
9	8	−1·23	7·12
	14	−0·65	1·98
	18	−0·48	1·07
	20	−0·42	0·82
10	9	−1·22	7·07
	16	−0·64	1·93
	21	−0·46	1·00
	24	−0·39	0·71
	25	−0·37	0·64

As n_{i1} and n_{i2} both become large, L_3 tends to $\sqrt{\{8/(N-1)\}}$, the third cumulant of the standardized chi-square distribution with $(N-1)$ degrees of freedom. To the extent that L_3 is close to this value, it is reasonably safe to assume that W is itself distributed as χ^2 with $(N-1)$ degrees of freedom. If, for example, $n_{i1} = n_{i2} = 10$ for all i, and $N = 10$, then $L_3 = 0{\cdot}82$, $\sqrt{\{8/(N-1)\}} = 0{\cdot}94$, and it is reasonably safe to assume that W is distributed as a chi-square variate with 9 degrees of freedom.

To facilitate the computation of κ_{4i} and κ_{6i}, Table 7 gives the values of these constants for all cases where $n_i \leqslant 10$. When $n_i \geqslant 10$, a satisfactory approximation to the value of κ_{6i} is given by

$$\kappa_{6i} = \frac{100}{21} \kappa_{4i}{}^2 .$$

Finally, Table 8 shows, as an example, the computation of z, W, and their significance tests, for a sample of 10 protocols. It can be seen that z is only significant at the 12·7 per cent level, and hence that the null hypothesis of stationarity is upheld for the overall group result. W is significant at the 1·4 per cent level, however, indicating significant heterogeneity of trends between the protocols. Hence, the evidence from this artificial example would not support a learning model predicting stationarity of individual protocols.

TABLE 8. EXAMPLE OF z AND W TESTS OF SIGNIFICANCE

Protocol	n_{1i}	n_{2i}	S_i	σ_i	x_i	κ_{4i}	κ_{6i}
1	4	6	16	9·38	1·71	−0·39	0·71
2	5	5	−11	9·57	−1·15	−0·37	0·64
3	10	4	−10	14·14	−0·71	−0·34	0·54
4	2	6	8	6·00	1·33	−0·67	2·07
5	7	6	22	14·00	1·57	−0·29	0·38
6	4	8	−24	11·78	−2·04	−0·36	0·60
7	10	4	12	14·14	0·85	−0·34	0·54
8	8	3	18	9·80	1·84	−0·45	0·95
9	9	4	−22	12·96	−1·70	−0·35	0·57
10	7	4	20	10·58	1·89	−0·37	0·64
					3·59	−3·92	7·64

$N = 10$

$$z_0 = \sum_{i=1}^{N} x_i / \sqrt{N} = 1{\cdot}14; \qquad Pr\{z \geqslant z_0\} = 0{\cdot}127.$$

$$W_0 = \sum_{i=1}^{N} x_i^2 - z^2 = 22{\cdot}43;$$

$$\pi_2 = 2(N-1) + \left(\frac{N-1}{N}\right)^2 \sum_{i=1}^{N} \kappa_{4i} = 14{\cdot}82;$$

$$\pi_3 = 8(N-1) + 12\left(\frac{N-1}{N}\right)^2 \sum_{i=1}^{N} \kappa_{4i} + \left(\frac{N-1}{N}\right)^3 \sum_{i=1}^{N} \kappa_{6i} = 39{\cdot}46;$$

$$\omega_0 = \frac{W_0 - (N-1)}{\sqrt{\pi_2}} = 3{\cdot}49;$$

$$L_3 = \pi_3 / \pi_2^{3/2} = 0{\cdot}691;$$

$$\xi_0 = \omega_0 - \frac{L_3}{6}\left(\omega_0^2 - 1\right) = 2{\cdot}20;$$

$$Pr\{\xi \geqslant \xi_0\} = 0{\cdot}014.$$

Acknowledgements

The research was carried out when the second author was at University College London. It was supported by a grant (MH 13950) from the National Institutes of Mental Health, U.S. Public Health Service.

References

Barton, D. E., David, F. N. and Mallows, C. L. (1958). Non-randomness in a sequence of two alternatives. *Biometrika* **45**, 166–177.

Binford, J. R. and Gettys, C. (1965). Nonstationarity in paired-associate learning as indicated by a second-guess procedure. *J. math. Psychol.* **2**, 190–195.

Bower, G. H. (1961). Application of a model to paired-associate learning. *Psychometrika* **26**, 255–280.

Bower, G. H. (1967). A descriptive theory of memory. In D. P. Kimble (Ed.), *The organization of recall: Proceedings of the second conference on learning, remembering & forgetting*, pp. 112–185. New York: Acad. Sci.

Bower, G. H. and Theios, J. (1964). A learning model for discrete performance levels. In R. C. Atkinson (Ed.), *Studies in Mathematical Psychology*. Stanford: Stanford University Press.

Cochran, W. G. (1950). The comparison of percentages in matched samples. *Biometrika* **37**, 256.

Cox, D. R. (1958). The regression analysis of binary sequences. *Jl. R. statist. Soc.* **20**, 216–242.

Hays, W. L. (1963). *Statistics for Psychologists*. New York: Holt, Rinehart and Winston.

Hilgard, E. R. (1938). A summary and evaluation of alternative procedures for the construction of Vincent curves. *Psychol. Bull.* **35**, 282–297.

Jonckheere, A. R. (1954 a). A distribution-free k-sample test against ordered alternatives. *Biometrika* **41**, 133–145.

Jonckheere, A. R. (1954 b). A test of significance for the relation between m rankings and k ordered categories. *Br. J. statist. Psychol.* **7**, 93–100.

Kendall, M. G. (1955). *Rank Correlation Methods*. London: Griffin and Co.

Kendall, M. G. (1958). *The Advanced Theory of Statistics*. Vol. I. London: Griffin and Co.

Luce, R. D. (1959). *Individual Choice Behaviour: a Theoretical Analysis*. New York: Wiley and Sons.

Restle, F. (1964). Sources of difficulty in learning paired-associates. In R. C. Atkinson (Ed.), *Studies in Mathematical Psychology*. Stanford: University Press.

Siegel, S. (1956). *Non-parametric Statistics for the Behavioural Sciences*. New York: McGraw-Hill.

Spence, K. W. (1956). *Behaviour Theory and Conditioning*. New Haven: Yale University Press.

Suppes, P. and Ginsberg, Rose (1963). A fundamental property of all-or-none models. *Psychol. Rev.* **70**, 139–161.

Vol. 20
Part 2
187–204

The British Journal of Mathematical and Statistical Psychology

November
1967

TYPE II ERRORS AND THEIR DECISION BASIS

By R. S. Rodger

University of Sydney

Methods are given for testing linear statistical hypotheses about population means of a variate which is normally distributed in each population, with common variance. The view is adopted that error-rates (or the expected proportions of decision errors) be prespecified for each decision to be made. This was the view that led to the invention of the *R* technique and methods for setting type II error-rates for this technique are given here. Experiment-based error-rates control the probability of making one or more decision errors. Although not an essential part of the doctrine, the decision-based error-rate view leads naturally to a consideration of the probability distributions of *r* decision errors and such distributions are given in this paper.

1. Introduction

It is now common practice to give some attention to error-rates of the first kind in the analysis of data from statistical experiments, though this is not always done explicitly. If the only error considered is the type I error, statements about the hypothesis accepted, when the null hypothesis is rejected, must be somewhat vague. In such a one-error context the sample size must be chosen on the rather arbitrary grounds of time and materials available for the experiment. Although such factors are not unimportant, at least of equal importance is the size of a treatment effect we hope to detect. A concern with the size of a treatment effect we intend a statistical experiment to detect involves consideration of alternatives to null hypotheses and error-rates of the second kind.

Given the choice of null hypotheses (${}_0\kappa_i$), at least one alternative to each (${}_1\kappa_i$), an error-rate of the first kind (α) for the ${}_0\kappa_i$, an error-rate of the second kind ($1-\beta$) for the ${}_1\kappa_i$ and a test procedure for the ${}_0\kappa_i$, the sample size (n) required to satisfy these conditions is determined. The specification of ${}_0\kappa_i$, ${}_1\kappa_i$, α, $1-\beta$ and test procedure not only makes explicit the intended amount of discrimination between hypotheses, it also compels proper emphasis on the design of statistical experiments.

In this paper we shall consider a populations in which the variate Y is normally distributed with means μ_j ($j=1, 2, \ldots, a$) and common variance σ_ϵ^2. It will be assumed that each of these populations is randomly sampled an equal number of times (n).

2. Null Hypotheses

The null hypotheses considered here are linear statistical hypotheses about the μ_j and the general form of the ith such hypothesis is:

$$ {}_0\kappa_i = c_{i1}\mu_1 + c_{i2}\mu_2 + \ldots + c_{ia}\mu_a + d_i = 0. \tag{1}$$

The c_{ij} are real numbers, not all zero, chosen by the investigator because of their relevance to the purpose of the investigation and d_i is a real number which may be zero. When, for fixed i, $\sum_j c_{ij}=0$, the hypothesis is termed a contrast. Any contrast in which one $c_{ij}=1$, another $c_{ij}=-1$ and all other $c_{ij}=0$, is a difference between two means (with or without the added d_i) and is termed a comparison. Examples of (1) for $a=4$ are

$$_0\kappa_1=\mu_1+\mu_2+\mu_3+\mu_4-100=0; \quad _0\kappa_2=\mu_1-\mu_2=0;$$
$$_0\kappa_3=\mu_3-\mu_4=0; \quad _0\kappa_4=0{\cdot}5(\mu_1+\mu_2)-0{\cdot}5(\mu_3+\mu_4)=0.$$

The values of c_{1j}, their sum of squares and d_1 are

$$c_{1j}=1,\ 1,\ 1,\ 1; \quad \sum_j c_{1j}^2=4; \quad d_1=-100.$$

This hypothesis is not a contrast because $\sum_j c_{1j}\neq 0$.

The last three hypotheses are contrasts, $_0\kappa_2$ and $_0\kappa_3$ being comparisons. Almost invariably reports of statistical experiments which use contrasts have all $d_i=0$, but there is no mathematical requirement that this be so.

Components of variation are quadratic functions of linear hypotheses and the test of a component will yield the same decision as the test of its corresponding hypothesis. The ith component is

$$_0\Lambda_i=n\Big(\sum_j c_{ij}\mu_j+d_i\Big)^2\Big/\sum_j c_{ij}^2=0. \tag{2}$$

The sample estimators of $_0\kappa_i$ and $_0\Lambda_i$ are

$$k_i=c_{i1}M_1+c_{i2}M_2+\ldots+c_{ia}M_a+d_i \tag{3}$$

and

$$L_i=n\Big(\sum_j c_{ij}M_j+d_i\Big)^2\Big/\sum_j c_{ij}^2, \tag{4}$$

in which the M_j are the sample means. If we multiply the c_{ij} and d_i of an hypothesis by some constant w_i, this will change the value of k_i for given M_j, but the value of L_i will be unchanged; hence the decision about $_0\Lambda_i$ and $_0\kappa_i$ is unaffected by such a multiplication. It is sometimes useful to consider *normalized* linear hypotheses. Such an hypothesis is one in which $\sum_j c_{ij}^2=1$. Any hypothesis may be converted to normalized form by using the multiplier $w_i=(\sum_j c_{ij}^2)^{-\frac{1}{2}}$.

If p hypotheses are chosen for a means and their c coefficients written as the matrix $_p\mathbf{C}_a$, the rank of this matrix cannot exceed a and when all p hypotheses are contrasts the rank of $_p\mathbf{C}_a$ cannot exceed $a-1$. The rank of $_p\mathbf{C}_a$ is the number of degrees of freedom available for estimating the p hypotheses.

Various interrelations are possible among hypotheses. The interrelations may be described by the matrix $_p\mathbf{C}_a\,{}_a\mathbf{C}_p{}^T$ and when this is diagonal (the non-zero elements being $\sum_j c_{ij}^2$), the hypotheses are mutually orthogonal. The index T indicates a transpose.

We may define ${}_p\mathbf{k}_1$ to have elements which are the sample values of p hypotheses with ${}_p\mathbf{C}_a$ of rank p. The total amount of variation among the M_j, for the given d_i, accounted for by these hypotheses is then

$$L_m = n\,{}_1\mathbf{k}_p^T({}_p\mathbf{C}_a\,{}_a\mathbf{C}_p^T)^{-1}\,{}_p\mathbf{k}_1. \tag{5}$$

If the hypotheses are mutually orthogonal, eqn. (5) simplifies to

$$L_m = \sum_i L_i. \tag{6}$$

If $p=a$ and all $d_i=0$, then eqn. (5) or (6) is the uncorrected sum of squares of the sample means, i.e.

$$L_m = \text{USSB} = n\sum_j M_j^2. \tag{7}$$

If $p=a-1$ contrasts and all $d_i=0$, then eqn. (5) or (6) is the component for the maximized contrast, which is the same as the corrected sum of squares between the sample means, as in analysis of variance, i.e.

$$L_m = \text{SSB} = n\sum_j (M_j - M.)^2, \tag{8}$$

where $M.$ is the general mean.

The traditional analysis of variance procedures test ${}_0\Lambda_m=0$, for which eqn. (8) is the estimator. Although this component plays an important part in statistical theory, it has been argued elsewhere (Rodger, 1967) that it is rarely of primary interest to empirical research workers. The component ${}_0\Lambda_m$ will be used in this paper mainly for theoretical purposes.

Relations (7) and (8) may be made to hold when all $d_i \neq 0$ by defining $M_j = \bar{Y}_j + m_j$. The $\bar{Y}_j$ are the observed means and they are adjusted by the m_j to allow for the d_i. The m_j are the elements of ${}_1\mathbf{m}_a$ given by

$${}_1\mathbf{m}_a = {}_1\mathbf{d}_p^T({}_p\mathbf{C}_a\,{}_a\mathbf{C}_p^T)^{-1}\,{}_p\mathbf{C}_a. \tag{9}$$

and ${}_1\mathbf{d}_p^T$ has elements d_i.

The decision rule that will be used in this paper is to reject those ${}_0\kappa_i$ for which

$$F_i = L_i/\nu_1\hat{\sigma}_\epsilon^2 \geq F_{(1-\alpha)}^{\nu_1} \tag{10}$$

and retain those which do not reach this criterion. In this expression $\hat{\sigma}_\epsilon^2$ is an estimate of σ_ϵ^2, with ν_2 degrees of freedom, and $F_{(1-\alpha)}^{\nu_1}$ is the value of the central variance ratio distribution, with ν_1 and ν_2 degrees of freedom, which has the area $(1-\alpha)^{\nu_1}$ to its left. This decision rule assumes that the test of ${}_0\kappa_i$ is double-ended, and this will be assumed in the rest of the paper. Should the test of ${}_0\kappa_i$ be single-ended, then α in (10) should be replaced by 2α. Tables of $F_{(1-\alpha)}^{\nu_1}$ are given in Rodger (1965).

3. ALTERNATIVE HYPOTHESES

An additive alternative to the general null hypothesis given by eqn. (1) is

$${}_1\kappa_i = c_{i1}\mu_1 + c_{i2}\mu_2 + \dots + c_{ia}\mu_a + d_i = \delta_i \tag{11}$$

and the quadratic form is

$$ {}_1\Lambda_i = n\delta_i^2 / \sum_j c_{ij}^2 = \sigma_\epsilon^2 \lambda_i^2. \tag{12} $$

The quantities δ_i and λ_i^2 are non-central parameters. The size of a treatment effect which will have a type II error-rate $(1-\beta)$ may be symbolized by $\lambda^2{}_{(1-\beta)}{}^{\nu_1}$.

Although the choice of δ_i is somewhat arbitrary (just as the choice of d_i, even $d_i = 0$, is arbitrary), as information about an empirical domain accumulates, the size of an effect it is thought desirable to detect (δ_i) can be ever better justified. The smaller δ_i is made, the larger must n be made to maintain constant error-rates of the first and second kind. In the absence of other information, a useful and not impractical choice is $\delta_i = \sigma_\epsilon \sqrt{\sum_j c_{ij}^2}$. Thus $\mu_1 = 0$ would be tested against $\mu_1 = \sigma_\epsilon$ and $\mu_1/\sqrt{2} - \mu_2/\sqrt{2} = 0$ against $\mu_1/\sqrt{(2-\mu_2)}/\sqrt{2} = \sigma_\epsilon$, which is the same as $\mu_1 - \mu_2 = \sigma_\epsilon \sqrt{2}$. As more is learned about populations, smaller values of δ_i may be used, if the scientific value of these smaller amounts justifies their cost in time, money and effort.

Examples of ${}_1\kappa_i$ alternatives to the ${}_0\kappa_i$ given in Section 2 are

$$ {}_1\kappa_1 = \mu_1 + \mu_2 + \mu_3 + \mu_4 - 100 = 2\sigma_\epsilon; $$

$$ {}_1\kappa_2 = \mu_1 - \mu_2 = \sigma_\epsilon \sqrt{2}; $$

$$ {}_1\kappa_3 = \mu_3 - \mu_4 = \sigma_\epsilon \sqrt{2}; $$

$$ {}_1\kappa_4 = 0{\cdot}5(\mu_1 + \mu_2) - 0{\cdot}5(\mu_3 + \mu_4) = \sigma_\epsilon. $$

These are the positive alternatives to ${}_0\kappa_i$. The negative alternatives are usually taken to be ${}_2\kappa_i = -\delta_i$. Such symmetry keeps $1-\beta$ for ${}_2\kappa_i$ the same as that for ${}_1\kappa_i$. In this paper the symbol ${}_1\kappa_i$ will often be used to refer to both alternatives to ${}_0\kappa_i$. It must be noted that the above alternatives are expressed in the unknown σ_ϵ units and there is no suggestion that any numerical value will be placed on this parameter.

In the above examples the δ_i differ, but each ${}_1\kappa_i$ is equally discriminable from its ${}_0\kappa_i$. The differences between the δ_i, in this case, are entirely due to differences in $\sum_j c_{ij}^2$. The fact that the size of δ_i may be affected by $\sum_j c_{ij}^2$ in this way makes the former a misleading non-central parameter. In this paper we prefer

$$ \lambda_i^2 = n\delta_i^2 / \sigma_\epsilon^2 \sum_j c_{ij}^2 \tag{13} $$

derived from eqn. (12). This is adapted from Patnaik (1949). If we apply eqn. (13) to the four above examples we obtain $\lambda_i^2 = n$ for all i. This is another way of expressing the useful rule that, in the absence of other information, we make $\delta_i = \sigma_\epsilon \sqrt{\sum_j c_{ij}^2}$.

Should we wish to consider p hypotheses, with coefficient matrix rank p, *simultaneously*, the non-central parameter for the alternatives to these, considered *simultaneously*, is

$$\lambda_m^2 = n_1 \boldsymbol{\delta}_p^T ({}_p\mathbf{C}_{a\,a}\mathbf{C}_p^T)^{-1}\,{}_p\boldsymbol{\delta}_1 \sigma_\epsilon^{-2} \tag{14}$$

in which the elements of ${}_p\boldsymbol{\delta}_1$ are the δ_i. If the p hypotheses are mutually orthogonal, then (14) reduces to

$$\lambda_m^2 = \sum_i \lambda_i^2 \tag{15}$$

and, should all $\lambda_i^2 = \lambda^2$ in such an orthogonal set, we have

$$\lambda_m^2 = p\lambda^2. \tag{16}$$

After testing $p = a$ null hypotheses, with coefficient matrix rank p, we may define ${}_p\bar{\boldsymbol{\delta}}_1$ to have zero elements for the ${}_0\kappa_i$ accepted, elements δ_i for those rejected in favour of ${}_1\kappa_i$ (because k_i is " significantly " positive) and $-\delta_i$ for those rejected in favour of ${}_2\kappa_i$ (because k_i is " significantly " negative). Such a set of decisions implies values for the μ_j, which are the elements of

$${}_1\boldsymbol{\mu}_a = {}_1\bar{\boldsymbol{\delta}}_p^T ({}_p\mathbf{C}_{a\,a}\mathbf{C}_p^T)^{-1}\,{}_p\mathbf{C}_a - {}_1\mathbf{m}_a, \tag{17}$$

${}_1\mathbf{m}_a$ being given by eqn. (9). If the hypotheses were $p = a-1$ contrasts, with coefficient matrix rank p, then the right-hand side of eqn. (17) yields the values of $\mu_j - \mu.$, where $\mu.$ is the general mean of the a populations.

Clearly if the rank of ${}_p\mathbf{C}_a$ is less than p, then ${}_p\mathbf{C}_{a\,a}\mathbf{C}_p^T$ is singular and its inverse does not exist. But a more important reason for insisting that decisions be restricted to p hypotheses with coefficient matrix rank p is that this avoids redundancy and contradiction among hypotheses. For example, in the statements $\mu_1 - \mu_2 = 0$, $\mu_2 - \mu_3 = 0$ and $\mu_1 - \mu_3 = 0$, $p = 3$ but ${}_p\mathbf{C}_a$ has rank 2. Each of these statements may be deduced from the other two; thus if the first two are true, then $\mu_1 = \mu_2 = \mu_3$ and the third follows by deduction. If one of the three contrasts above is said to be false (whether δ_i is stated or not), the other two cannot both be true. Thus, if the first is said to be false and the other two true this is a contradiction because the two true hypotheses imply that $\mu_2 = \mu_3 = \mu_1$, yet it is claimed for the first that $\mu_1 \neq \mu_2$. In the same way, if we state that $\mu_1 - \mu_2 = \delta_1$, $\mu_2 - \mu_3 = \delta_2$ and $\mu_1 - \mu_3 = \delta_3$ are true, this contains a redundant statement when $\delta_3 = \delta_1 + \delta_2$ and a contradiction when $\delta_3 \neq \delta_1 + \delta_2$.

The requirement that δ_i be expressed in σ_ϵ units is, on occasion, too vague and we wish to express alternatives to the ${}_0\kappa_i$ in the measured units of Y. Stein (1945) has shown that this is possible if we adopt a two-stage sampling plan. With this procedure we specify α, ${}_0\kappa_i$, $1-\beta$, ${}_1\kappa_i$ and ${}_2\kappa_i$, draw n replicates and use the resulting data to calculate $\hat{\sigma}_\epsilon^2$ with ν_2 degrees of freedom. The non-central parameter for hypothesis ${}_1\kappa_i$ will be

$$\lambda_i'^2 = N\delta_i'^2 / \hat{\sigma}_\epsilon^2 \sum_j c_{ij}^2, \tag{18}$$

if we use a full N replicates. The symbol δ_i' is used when ${}_1\kappa_i$ is expressed in the measured units of Y to distinguish it from δ_i which is expressed in σ_ϵ units.

Given ν_1, ν_2, $\hat{\sigma}_\epsilon^2$, ${}_0\kappa_i$, α, ${}_1\kappa_i$ and $1-\beta$, we may calculate the N required, in the light of first stage data, to be

$$N=[{\lambda'}^2_{(1-\beta)}{}^{\nu_1}\hat{\sigma}_\epsilon^2\sum_j c_{ij}^2/\delta_i'^2]+1. \tag{19}$$

The square brackets indicate that the fractional part is truncated. If $N\leqslant n$, we now test our hypotheses in the usual way, by putting eqn. (4) for L_i in eqn. (10), otherwise we draw $N-n$ further replicates and put

$$L_i=N(\sum_j c_{ij}M_j+d_i)^2/\sum_j c_{ij}^2 \tag{20}$$

in eqn. (10) with $\hat{\sigma}_\epsilon^2$ and ν_2 being the values found after the first stage of sampling and the M_j being the means of the N replicates. This procedure does not make the power of the tests *strictly* independent of σ_ϵ^2, but it is a slight modification of one which does do so (i.e. if $N\leqslant n$ in eqn. (19) make $N=n+1$) and it avoids wasting information (Stein, 1945, p. 247).

The fact that δ_i' involves no nuisance parameter is an obvious advantage, but a disadvantage of two-stage sampling is that, unless caution is used, the N found after the first stage may be impractically large. The size of N depends on the size of n, among other things, and it is possible to develop methods for optimizing the choice of n (Seelbinder, 1953, and Moshman, 1958). Here we note only that the probability is ω of finding

$$\nu_2\hat{\sigma}_\epsilon^2/\sigma_\epsilon^2\geqslant\chi^2_{1-\omega}; \tag{21}$$

so the probability is ω of finding

$$\hat{\sigma}_\epsilon^2\geqslant\sigma_\epsilon^2F_{1-\omega}, \tag{22}$$

in which $F_{1-\omega}$ has degrees of freedom ν_2 in the numerator and ∞ in the denominator. Substituting the right-hand side of eqn. (22) in eqn. (19), we have the probability ω that

$$N\geqslant[{\lambda'}^2_{(1-\beta)}{}^{\nu_1}\sigma_\epsilon^2F_{1-\omega}\sum_j c_{ij}^2/\delta_i'^2]+1. \tag{23}$$

This relation may be used to compute the likely upper value of N for any n, given a reasonable choice of σ_ϵ^2.

Examples of ${}_1\kappa_i=\delta_i'$ alternatives to the ${}_0\kappa_i$ given in Section 2 are

$$\begin{aligned}
{}_1\kappa_1&=\mu_1+\mu_2+\mu_3+\mu_4-100=7;\\
{}_1\kappa_2&=\mu_1-\mu_2=\sqrt{24\cdot5};\\
{}_1\kappa_3&=\mu_3-\mu_4=\sqrt{24\cdot5};\\
{}_1\kappa_4&=0\cdot5(\mu_1+\mu_2)-0\cdot5(\mu_3+\mu_4)=3\cdot5.
\end{aligned}$$

Each of these ${}_1\kappa_i$ has $\lambda_i'^2=12\cdot25N/\hat{\sigma}_\epsilon^2$. If all ${}_0\kappa_i$ used a common $\hat{\sigma}_\epsilon^2$, then all ${}_1\kappa_i$ are equally discriminable from their ${}_0\kappa_i$ partners.

For a null hypothesis which is false there is no risk of a type I error; the risk is that we fail to accept the true alternative—a type II error. Some statisticians have written of type III errors, e.g. accepting ${}_2\kappa_i$ when ${}_1\kappa_i$ is true, but such an event will be sufficiently rare to be of no practical importance when $1-\beta$ is small. When ${}_0\kappa_i$ is false the sampling distribution of F_i in eqn. (10) is a non-central variance ratio distribution. The non-central distribution appropriate to one-stage sampling (δ_i expressed in σ_ϵ units) was found by Fisher (1928) and studied by Wishart (1932), Tang (1938) and Patnaik (1949). This distribution has parameters ν_1, ν_2 and, for the methods used in this paper, a non-central parameter $\nu_1\lambda^2_{(1-\beta)}{}^{\nu_1}$. The non-central distribution which applies in two-stage sampling (δ_i' expressed in the measured units of Y) was given by Stein (1945, p. 251, eqn. 34). This has parameters ν_1, ν_2 and $\nu_1\lambda'^2_{(1-\beta)}{}^{\nu_1}$.

If we symbolize the type II error-rate by $1-\beta$, the area to the left of $F_{(1-\alpha)}{}^{\nu_1}$ in either non-central distribution is $(1-\beta)^{\nu_1}$.

4. PRESCRIBED TESTS

When $a=1$, we may choose c_{11} and d_1, independently of the data to be used for testing, and test

$${}_0\kappa_1=c_{11}\mu_1+d_1=0 \text{ against } {}_1\kappa_1=c_{11}\mu_1+d_1=\delta_1 \text{ and } {}_2\kappa_1=c_{11}\mu_1+d_1=-\delta_1$$

by using eqns. (4) and (10) with $\nu_1=1$ and $\nu_2=n-1$. The degrees of freedom ν_2 are for the mean-square within the sample, which is used as the estimator $\hat{\sigma}_\epsilon^2$. The sampling distribution of F_i in eqn. (10) is the variance ratio distribution with $\nu_1=1$, $\nu_2=n-1$ and $\nu_1\lambda^2_{(1-\beta)}{}^{\nu_1}=0$ (the central case) when ${}_0\kappa_1$ is true and $\nu_1\lambda^2_{(1-\beta)}{}^{\nu_1}=1\times n\delta_1^2/\sigma_\epsilon^2\sum_j c_{ij}^2$ (the non-central case) when either ${}_1\kappa_1$ or ${}_2\kappa_1$ is true. This test is equivalent to that termed "the one-sample t test" in elementary text books. If p independent investigations of this kind are done, each with error-rates α and $1-\beta$, the probability of rejecting r of the null hypotheses is the $(r+1)$th term in the expansion of $(\gamma+\alpha)^p$, when all null hypotheses are true and in the expansion of $(\theta+\beta)^p$ when an alternative hypothesis is always true ($\gamma=1-\alpha$ and $\theta=1-\beta$).

Frequently the hypotheses of interest in empirical science involve a populations, with $a>1$. In such cases we may choose c_{1j} and d_1, independently of the data to be used for testing, and test eqn. (1) against eqn. (11) by using eqn. (4) substituted in (10). The sampling distributions of F_i in eqn. (10), the error-rates and the probabilities of r rejections of ${}_0\kappa_1$ in p independent tests are the same as those above, except that we would now use the within samples mean-square (MSW) as the estimator of σ_ϵ^2, as in analysis of variance, with $\nu_2=a(n-1)$. An application of this procedure, which is in common use, is the so-called "two-sample t test", in which $a=2$, but we test only one hypothesis (${}_0\kappa_1=\mu_1-\mu_2=0$) against its alternatives.

If p mutually orthogonal hypotheses are chosen in advance to be tested by using one set of data (n random replicates from each of $a\geqslant p$ populations), we may make the tests independent by specifying in advance p independent estimators of σ_ϵ^2, one for each hypothesis. Many independent estimators of

σ_ϵ^2 are possible in such data, the a within sample mean-squares ($\hat{\sigma}_i^2$), each with $\nu_2=n-1$, will satisfy this requirement. When the data are collected, each hypothesis will be tested by putting eqn. (4) for L_i in eqn. (10), with $\hat{\sigma}_\epsilon^2$ replaced by $\hat{\sigma}_i^2$, $\nu_1=1$ and $\nu_2=n-1$. The sampling distributions of F_i, the error-rates and the probabilities of r rejections of ${}_0\kappa_i$ are those given in the first paragraph of this section.

Since the power (β) of the eqn. (10) test of eqn. (1) against eqn. (11) diminishes rapidly (for small ν_2) as ν_2 decreases and the $\hat{\sigma}_i^2$ are all estimates of the same parameter (σ_ϵ^2), it is desirable to use, in eqn. (10), the single, common estimator of σ_ϵ^2, which is MSW with $\nu_2=a(n-1)$. The use of a common estimator (MSW) may well increase the robustness of the tests and it will simplify the application of two-stage sampling. When MSW is used the tests are no longer independent and the distribution of r rejections of ${}_0\kappa_i$ is, therefore, no longer Bernoullian binomial (though this is not an unreasonable, practical approximation to the distribution). Because $\sum_i L_i=L_m$ for mutually orthogonal hypotheses, and the F_i have a common denominator ($1\times$ MSW), then $F_m=\sum_i F_i/p$ is distributed in variance ratio form with $\nu_1=p$, $\nu_2=a(n-1)$ and $\lambda_m^2=\sum_i \lambda_i^2$ (for those ${}_0\kappa_i$ which are true, $\lambda_i^2=0$). When $\nu_1=1$, then $F_{(1-\alpha)}{}^{\nu_1}=t^2_{1-\frac{1}{2}\alpha}$. If it is anticipated that r or more of the ${}_0\kappa_i$ will be rejected, then F_m will be at least $rp^{-1}t^2_{1-\frac{1}{2}\alpha}$; so the *maximum* probability of rejecting *r or more* of the ${}_0\kappa_i$ is the area to the right of $rp^{-1}t^2_{1-\frac{1}{2}\alpha}$ in the aforementioned variance ratio distribution with $\nu_1=p$. If p intercorrelated hypotheses, with ${}_p\mathbf{C}_a$ of rank p, were chosen, they may be tested by the above procedure. The type I and type II error-rates in each test, considered separately, will be α and $1-\beta$, but the maximum probability of rejecting r or more of the ${}_0\kappa_i$ now depends on the nature of the intercorrelation of hypotheses. For any given ${}_p\mathbf{C}_a\,{}_a\mathbf{C}_p{}^T$, the different possible selections of r of the ${}_0\kappa_i$ may yield different probability maxima. We may compute

$$
{}_1\mathbf{t}_p{}^T\,{}_p\mathbf{C}_a\,{}_a\mathbf{C}_p{}^T\,{}_p\mathbf{t}_1\,p^{-1} \tag{24}
$$

for each of the possible selections of r from p. The elements of ${}_p\mathbf{t}_1$ are $(\sum_j c_{ij}^2)^{-\frac{1}{2}}t_{1-\frac{1}{2}\alpha}$ for those r of the ${}_0\kappa_i$ selected and zero for the others. We define F to be the smallest of the $\binom{p}{r}$ values of eqn. (24). The *maximum* probability of rejecting r or more of the ${}_0\kappa_i$ is the area to the right of F in the variance ratio distribution with $\nu_1=p$, $\nu_2=a(n-1)$ and $\lambda_m^2=0$ when all ${}_0\kappa_i$ are true, and λ_m^2 given by eqn. (14) when all ${}_1\kappa_i$ are true.

The non-central parameter $\lambda^2_{(1-\beta)}{}^{\nu_1}$ for alternatives to ${}_0\kappa_i$ can be found by numerical integration of the classical non-central variance ratio distribution. These values, for $\nu_1=1\times 1$ to 8, $\alpha=1-\beta=0{\cdot}05$ and various ν_2 are given in Table 1. A more extensive table of these quantities can be obtained from the author.

TABLE 1. VALUES OF $\lambda^2_{0\cdot05\nu_1}$ FOR $\alpha=0\cdot05$

ν_2	$\nu_1=1$	2	3	4	5	6	7	8
4	24·285	23·963	24·420	25·048	25·708	26·356	26·980	27·575
6	19·067	18·258	18·303	18·593	18·974	19·390	19·816	20·242
8	17·112	16·096	15·934	16·036	16·249	16·515	16·807	17·111
10	16·101	14·978	14·699	14·690	14·799	14·970	15·174	15·397
12	15·486	14·299	13·946	13·865	13·906	14·012	14·155	14·321
14	15·074	13·844	13·441	13·310	13·303	13·363	13·463	13·587
16	14·778	13·518	13·079	12·912	12·870	12·896	12·962	13·056
18	14·556	13·274	12·808	12·613	12·544	12·544	12·585	12·653
20	14·383	13·084	12·597	12·381	12·290	12·269	12·290	12·338
30	13·887	12·542	11·995	11·717	11·565	11·483	11·444	11·435
40	13·652	12·285	11·712	11·404	11·223	11·112	11·045	11·006
50	13·515	12·136	11·546	11·222	11·024	10·895	10·811	10·757
60	13·425	12·039	11·438	11·103	10·894	10·754	10·659	10·594
∞	12·995	11·572	10·924	10·537	10·275	10·083	9·936	9·818

If p mutually orthogonal hypotheses are pre-specified to be tested by two-stage sampling, each with $\nu_1=1$, and p independent estimators of σ_ϵ^2 (say $\hat{\sigma}_i^2$) are pre-assigned—one to each hypothesis, then the type I error-rate for each true ${}_0\kappa_i$ tested by substituting eqn. (20) in eqn. (10) will be α. The probability of r type I errors in p true null hypotheses will be the $(r+1)$th term in the expansion of $(\gamma+\alpha)^p$. Stein's modified rule for N ensures that the type II error-rate for each true alternative will not exceed $1-\beta$, hence the *maximum* probability of r *or more* type II errors in p true alternatives is the sum of the first $(r+1)$ terms in the expansion of $(\theta+\beta)^p$.

The use of $\hat{\sigma}_i^2$ in eqn. (19) will generally yield a different N_i for each hypothesis. Such different N_i will generally make hypotheses, which were orthogonal at the first stage, intercorrelated after the second stage. For this reason, as well as those given earlier in this section, the use of MSW is recommended. If this is done the *maximum* probability of r *or more* rejections of ${}_0\kappa_i$ is the area to the right of $rp^{-1}t^2_{1-\frac{1}{2}\alpha}$ in the central or non-central variance ratio distribution with $\nu_1=p$, $\nu_2=a(n-1)$ and $\lambda_m'^2=0$ when all ${}_0\kappa_i$ are true and $\lambda_m'^2=\sum_i \lambda_i'^2$ when an alternative hypothesis is always true. In this case the non-central distribution is Stein's special form.

The distribution of r rejections of ${}_0\kappa_i$ with p pre-specified, intercorrelated hypotheses can be handled in the same way as the one-stage sampling plan.

As an example of the application of the prescribed tests strategy, consider an investigation into the relative effectiveness of different methods of teaching elementary statistics. The four procedures to be studied are:

(1) the use of a traditional type of textbook and lectures,

(2) the same textbook, but tutorials in place of lectures,

(3) the use of a branched-program textbook, and

(4) the use of a linear-program textbook.

Students' knowledge of statistics will be assessed by a multiple-choice test administered after nine weeks of study. We specify

$$
\begin{array}{lll}
{}_0\kappa_1=\mu_1-\mu_2=0, & {}_1\kappa_1=\mu_1-\mu_2=\sigma_\epsilon\sqrt{2}, & {}_2\kappa_1=\mu_1-\mu_2=-\sigma_\epsilon\sqrt{2}, \\
{}_0\kappa_2=\mu_3-\mu_4=0, & {}_1\kappa_2=\mu_3-\mu_4=\sigma_\epsilon\sqrt{2}, & {}_2\kappa_2=\mu_3-\mu_4=-\sigma_\epsilon\sqrt{2}, \\
{}_0\kappa_3=\mu_1+\mu_2-\mu_3-\mu_4=0, & {}_1\kappa_3=\mu_1+\mu_2-\mu_3-\mu_4=2\sigma_\epsilon, & {}_2\kappa_3=\mu_1+\mu_2-\mu_3-\mu_4 \\
 & & \qquad =-2\sigma_\epsilon,
\end{array}
$$

and set $\alpha=1-\beta=0{\cdot}05$ for each of these.

If we use MSW with $\nu_2=a(n-1)$ to test all ${}_0\kappa_i$, then the various ν_2 in Table 1 are equivalent to $n=2$ by steps of 0·5 to 6, 8·5, 11, 13·5, 16 and ∞. All alternative hypotheses have $\lambda_i^2=n$; so we must choose n to satisfy $n\geqslant\lambda^2_{0\cdot05}$. Clearly $n=14$, giving $\nu_2=4\times13=52$, satisfies this requirement.

Students are now assigned at random to the four treatment groups and, after nine weeks of study, their attainment assessed. It is found that MSW = 10, $M_j=11, 14, 12, 10$, giving $L_i=63, 28, 31{\cdot}5$ and $F_i=6{\cdot}3, 2{\cdot}8$ and $3{\cdot}5$ ($j=1, 2, 3, 4$; $i=1, 2, 3$). Only the first of these reaches $F_{0\cdot95}=4{\cdot}03$; so ${}_0\kappa_1$ is rejected in favour of ${}_2\kappa_1$ and the remaining ${}_0\kappa_i$ retained.

The values of μ_j implied by these decisions can be found by inserting ${}_1\bar{\boldsymbol{\delta}}_3{}^T=(-\sigma_\epsilon\sqrt{2}\quad 0\quad 0)$ in eqn. (17) which gives

$$\mu_j=\mu.-\sigma_\epsilon/\sqrt{2},\quad \mu.+\sigma_\epsilon/\sqrt{2},\quad \mu.,\quad \mu.$$

The maximum probability of $r=1$ or more type I errors, if all three ${}_0\kappa_i$ were true, is the area to the right of $rp^{-1}\,t^2_{1-\frac{1}{2}\alpha}=1\times4{\cdot}03/3=1{\cdot}343$ in the variance ratio distribution with $\nu_1=p=3$, $\nu_2=52$ and $\lambda_m{}^2=0$, which is 0·270 (the binomial distribution gives 0·143 as the probability of rejecting $r=1$ or more of the ${}_0\kappa_i$).

As an example of two-stage sampling an experiment is designed to investigate the effects of a placebo (X_1) and varying amounts of the drug magnesium pemoline (doses 10, 20, 30 and 40 mg/kg being X_2 to X_5) on the discrimination learning of rats, thirty minutes after intra-peritoneal injection. It is planned that linear, quadratic, cubic and quartic components of trend will be tested. These are the regression components of μ_j (the mean learning score of population j) on $x_j=(X_j-20)/10$ and other polynomials in x_j which are, in this case, $x_j{}^2-2$, $5x_j{}^3-17x_j$ and $7x_j{}^4-31x_j{}^2+14{\cdot}4$. These polynomials yield, apart from scaling factors, the c_{ij} (see Kendall and Stuart, 1961). The null hypotheses and their alternatives are

$$
\begin{array}{ll}
{}_0\kappa_1=-2\mu_1-\mu_2+\mu_4+2\mu_5=0, & {}_1\kappa_1=-2\mu_1-\mu_2+\mu_4+2\mu_5=20, \\
{}_0\kappa_2=2\mu_1-\mu_2-2\mu_3-\mu_4+2\mu_5=0, & {}_1\kappa_2=2\mu_1-\mu_2-2\mu_3-\mu_4+2\mu_5=\sqrt{560}, \\
{}_0\kappa_3=-\mu_1+2\mu_2-2\mu_4+\mu_5=0, & {}_1\kappa_3=-\mu_1+2\mu_2-2\mu_4+\mu_5=20, \\
{}_0\kappa_4=\mu_1-4\mu_2+6\mu_3-4\mu_4+\mu_5=0, & {}_1\kappa_4=\mu_1-4\mu_2+6\mu_3-4\mu_4+\mu_5=10\sqrt{28},
\end{array}
$$

$$
\begin{array}{l}
{}_2\kappa_1=-2\mu_1-\mu_2+\mu_4+2\mu_5=-20, \\
{}_2\kappa_2=2\mu_1-\mu_2-2\mu_3-\mu_4+2\mu_5=-\sqrt{560}, \\
{}_2\kappa_3=-\mu_1+2\mu_2-2\mu_4+\mu_5=-20, \\
{}_2\kappa_4=\mu_1-4\mu_2+6\mu_3-4\mu_4+\mu_5=-10\sqrt{28}.
\end{array}
$$

We set $\alpha=1-\beta=0{\cdot}05$.

When $\nu_1 = 1$, Stein's non-central parameter has the value

$$\lambda'^2_{(1-\beta)} = (t_{1-\frac{1}{2}\alpha} + t_\beta)^2, \tag{25}$$

in which t is Student's statistic with ν_2 degrees of freedom.

If we used $n = 9$, then $\nu_2 = a(n-1) = 40$ for MSW and $\lambda'^2_{0\cdot05} = (2\cdot021 + 1\cdot684)^2 = 13\cdot727$. We might have some preliminary evidence that MSW is not likely to exceed 50. Since $\lambda_i'^2 = 40N/\text{MSW}$ for all the above hypotheses, then from eqn. (19) we obtain $N = [13\cdot727 \times 50/40] + 1 = 18$. If we insert $\sigma_\epsilon^2 = 50$ in eqn. (23) with $\omega = 0\cdot05$, the probability is 0·05 that $N \geqslant [13\cdot727 \times 50 \times 1\cdot69/40] + 1 = 29$. These N's are considered practicable; so $n = 9$ is chosen for the first stage of sampling. Note that if n were increased then, all other things equal, the expected value of N would decrease, but there is a risk of overshooting the mark and finding $N < n$ at the end of the first stage. Increasing n too far would be uneconomic.

After the first stage of sampling $an = 45$ animals, the experiment is run and the value MSW $= 35$ obtained from an analysis of the learning scores. From eqn. (19), $N = [13\cdot727 \times 35/40] + 1 = 13$ is computed. Four more replicates are now drawn for each of the $a = 5$ treatments, these animals treated in exactly the same way as their predecessors and the observed mean learning scores for the $N = 13$ replicates from each population were $M_j = 20\frac{5}{13}, 22, 22\frac{10}{13}, 23, 27\frac{5}{13}$, eqn. (20) gives $L_i = 292\cdot5, 23\cdot2, 32\cdot5, 3\cdot79$ and eqn. (10) gives $F_i = 8\cdot36, 0\cdot66, 0\cdot93$ and $0\cdot11$ $(j = 1, 2, 3, 4, 5;\ i = 1, 2, 3, 4)$. Only the first of these is larger than $F_{0\cdot95} = 4\cdot08$; so ${}_0\kappa_1$ is rejected in favour of ${}_1\kappa_1$ and the other ${}_0\kappa_i$ retained.

The values of μ_j implied by these decisions can be found by inserting ${}_1\bar{\boldsymbol{\delta}}_4{}^T = (20\ 0\ 0\ 0)$ in eqn. (17), which gives

$$\mu_j = \mu. - 4,\quad \mu. - 2,\quad \mu.,\quad \mu. + 2,\quad \mu. + 4.$$

The maximum probability of $r = 1$ or more type I errors if all ${}_0\kappa_i$ were true is the area to the right of $rp^{-1}t^2_{1-\frac{1}{2}\alpha} = 1 \times 4\cdot08/4 = 1\cdot02$ in the variance ratio distribution with $\nu_1 = p = 4$, $\nu_2 = 40$ and $\lambda_m{}^2 = 0$, which is 0·409 (the binomial distribution gives 0·185).

It should be noted that the value claimed for an hypothesis when ${}_0\kappa_i$ is rejected may be much larger than the minimum observed amount necessary for rejection of ${}_0\kappa_i$. The minimum required to reject ${}_0\kappa_1$ in favour of ${}_1\kappa_1$ in the two-stage sampling illustration is $k_1 = 10\cdot42$, yet the amount claimed is ${}_1\kappa_1 = 20$. The difference between the minimum amount required for rejection and the amount accepted depends on $1-\beta$, among other things. The smaller $1-\beta$ is made, the larger does this difference become and the "price" of this is the increase in N. When $1-\beta = 0\cdot5$ there is no difference in these amounts, apart from a truncation factor. Analogous relations hold in the one-stage sampling case, but they are complicated by the presence of the unknown σ_ϵ in δ_i.

It is interesting to note that there is enough variation among the M_j in the one-stage sampling example to reject three null hypotheses. Had we chosen a different set of hypotheses, we might have found $L_1=40{\cdot}8$, $L_2=40{\cdot}8$ and $L_3=40{\cdot}9$. These account for the same M_j variation as the L_i we found, because $\mathrm{SSB}=\sum_i L_i=122{\cdot}5$. Our hypothetical set of L_i give $F_i=4{\cdot}08$, $4{\cdot}08$, $4{\cdot}09$ and all three of the null hypotheses would have been rejected. In the two-stage sampling illustration the variation among the M_j could support the rejection of two appropriate ${}_0\kappa_i$ chosen in advance. The choice of ${}_0\kappa_i$ that was actually made has over-concentrated " significance " in ${}_0\kappa_1$ and there is a sense in which this choice was unwise. In the formalities of decision, making a large difference between F_i and $F_{(1-\alpha)\nu_1}$ has the same effect as a small difference and, for this reason, such large differences waste information. This is the strongest statistical justification for *post hoc* data snooping. But if we wish to test statistical hypotheses with the advantage of hindsight, the sampling distributions which are appropriate are not those used for testing in this section. To this problem we now turn.

5. Unrestricted Tests

In this section we assume all $d_i=0$. This does not prevent us from pre-specifying values of m_j and carrying out an analysis of $M_j=\bar{Y}_j+m_j$ by the methods of this section—the $\bar{Y}_j$ being the observed sample means.

An infinite number of different sets of p null hypotheses, with ${}_p\mathbf{C}_a$ having rank p, are possible for a population means. Here we elaborate a method given by Rodger (1967) for testing p hypotheses, the choice of which is influenced by the observed M_j. There are two criticisms that can be made of testing hypotheses suggested by the data used to test them. The first is a statistical criticism which claims that a procedure of this kind capitalizes on sampling error. We take this to mean that the α and $1-\beta$ claimed are false, the former, at least, being a gross under-statement of the true type I error-rate, if we seek ${}_0\kappa_i$ to reject. The procedures of this section are not open to this criticism, since the *actual* error-rates (or the expected proportions of type I and II errors) are the α and $1-\beta$ pre-specified. The second criticism is scientific and says that a set of statements (decisions) without adequate rationale is of little, if any, scientific value. But when the choice of hypotheses is influenced by the observed M_j it is not inevitable that they must have an inadequate rationale. Frequently the research worker has a good rationale for more than one set of p hypotheses with ${}_p\mathbf{C}_a$ of rank p and he has difficulty in selecting which set to test by the methods of Section 4. It is also true that exploration plays a fundamental role in empirical science.

The method elaborated in this section should not be interpreted as implying that the author supports the proliferation of publications of scientific findings which are given little rationale.

If no restrictions were placed on the kind and number of hypotheses which might be chosen *post hoc* for formal decision making, it would be possible to state only very crude bounds for the error-rates α and $1-\beta$. We shall adopt the rule that as many ${}_0\kappa_i$ as possible be rejected and, initially, restrict the decisions to be made to p mutually orthogonal hypotheses. If these ${}_0\kappa_i$ have at least one $\sum c_{ij} \neq 0$, we make $p = a$, otherwise, for contrasts only, we make $p = a - 1$.

Consider independent tests of ${}_0\Lambda_i = 0$ with a decision rule which ensures that α and $1 - \beta$ are the error-rates for ${}_0\kappa_i$ and ${}_1\kappa_i$ respectively. Under these conditions the probability of retaining all ${}_0\kappa_i$ is $(1-\alpha)^p$ when all ${}_0\kappa_i$ are true and $(1-\beta)^p$ when all ${}_1\kappa_i$ are true. The retention of all ${}_0\kappa_i$ implies the retention of ${}_0\Lambda_m = 0$. The probability of retaining ${}_0\Lambda_m = 0$ when it is true will be $(1-\alpha)^p$ if we reject it when

$$F_m = L_m/\nu_1 \text{MSW} \geqslant F_{(1-\alpha)^{\nu_1}} \tag{26}$$

with $\nu_1 = p$. But when the L_i are derived from mutually orthogonal hypotheses and the F_i in eqn. (10) uses MSW for $\hat{\sigma}_\epsilon^2$, then

$$F_m = \sum_i L_i/\nu_1 \text{MSW} = \sum_i F_i. \tag{27}$$

The use of a common error-variance estimate is advisable for the reasons given for prescribed tests in Section 4, but its use makes the tests no longer independent.

The use of L_i, chosen *post hoc*, in eqn. (10) with $\hat{\sigma}_\epsilon^2 = \text{MSW}$ and $\nu_1 = p$ is the R technique introduced by Rodger (1965 and 1967). We maximize the number of ${}_0\kappa_i$ rejected when the number rejected is

$$q = [F_m/F_{(1-\alpha)^{\nu_1}}], \tag{28}$$

the square brackets indicating truncation of the fractional part. If eqn. (28) yields $q > \nu_1$, we make $q = \nu_1$.

The probability of rejecting r null hypotheses is the area between $rF_{(1-\alpha)^{\nu_1}}$ and $(r+1)F_{(1-\alpha)^{\nu_1}}$ in the variance ratio distribution with $\nu_1 = p$, $\nu_2 = a(n-1)$ and $\lambda_m{}^2 = 0$ when all ${}_0\kappa_i$ are true and $\lambda_m{}^2 = \sum_i \lambda_i{}^2$ when some of the p alternatives are true (for $r = \nu_1$ we make $(r+1)F_{(1-\alpha)^{\nu_1}} = \infty$). If the resulting probabilities are symbolized by $P_0, P_1, \ldots, P_{\nu_1}$, the expected proportion of type I errors in $p = \nu_1$ true ${}_0\kappa_i$ tested in a single application of the R method is

$$E(\alpha) = \sum_{r=0}^{\nu_1} rP_r/\nu_1. \tag{29}$$

The discrepancy between α and $E(\alpha)$ indicates one of the effects of non-independence of the tests (the other effect being the form of distribution of r rejections of ${}_0\kappa_i$). In the cases investigated by numerical integration so far ($\nu_1 = 2$ to 20, $\nu_2 = 4$ to ∞ and $\alpha = 0{\cdot}05$), α is a good approximation to $E(\alpha)$—$E(\alpha)$ ranging from 0·078 to 0·033 when $\nu_1 \leqslant \nu_2$—certainly quite satisfactory for practical purposes.

If the alternatives to all ${}_0\kappa_i$ have an equal value of λ^2 and all ${}_1\kappa_i$ are true, the probability will be $(1-\beta)^p$ of retaining ${}_0\Lambda_m = 0$ (hence all ${}_0\kappa_i$) when the area to the left of $F_{(1-\alpha)^{\nu_1}}$ is $(1-\beta)^{\nu_1}$ in the variance ratio distribution with $\nu_1 = p$, $\nu_2 = a(n-1)$ and $\lambda_m{}^2 = \nu_1 \lambda^2{}_{(1-\beta)^{\nu_1}}$. The successive areas between $rF_{(1-\alpha)^{\nu_1}}$ and $(r+1)F_{(1-\alpha)^{\nu_1}}$ ($r = 0, 1, \ldots, \nu_1$; and $(r+1)F_{(1-\alpha)^{\nu_1}} = \infty$ when $r = \nu_1$) in this distribution are the probabilities of rejecting r of the ${}_0\kappa_i$ when all ${}_1\kappa_i$ are true and these may be symbolized by $Q_0, Q_1, \ldots, Q_{\nu_1}$. The expected proportion of rejections of ${}_0\kappa_i$ in a single application of the R technique in this case is

$$E(\beta) = \sum_{r=0}^{\nu_1} rQ_r/\nu_1. \tag{30}$$

Numerical integration of the distributions with the ν_1, ν_2 and $\lambda^2_{0\cdot 05^{\nu_1}}$ given in Table 1 yield $E(\beta)$ ranging from 0·984 to 0·937 ($\beta = 0{\cdot}95$).

We may choose F_{π_0} in place of $F_{(1-\alpha)^{\nu_1}}$ as a critical value and symbolize the successive areas between rF_{π_0} and $(r+1)F_{\pi_0}$ in the variance ratio distribution with $\nu_1 = p$, $\nu_2 = a(n-1)$ and $\lambda_m{}^2 = 0$ by $\pi_0, \pi_1, \ldots, \pi_{\nu_1}$. An appropriate choice of F_{π_0} will yield

$$E(\alpha) = \sum_{r=0}^{\nu_1} r\pi_r/\nu_1 = \alpha. \tag{31}$$

The corresponding areas in the variance ratio distribution with $\nu_1 = p$, $\nu_2 = a(n-1)$ and $\lambda_m{}^2 = \nu_1\lambda^2_{\phi_0}$ may be symbolized as $\phi_0, \phi_1, \ldots, \phi_{\nu_1}$. An appropriate choice of $\lambda_{\phi_0}{}^2$ will yield

$$E(\beta) = \sum_{r=0}^{\nu_1} r\phi_r/\nu_1 = \beta. \tag{32}$$

Values of F_{π_0} and $\lambda_{\phi_0}{}^2$ that satisfy eqns. (31) and (32) for popular values of α and β are currently being sought.

Stein (1945) discussed the use of his special form of non-central variance ratio distribution when $\nu_1 > 1$ and Healy (1956) showed how it could be used with two-stage sampling for simultaneous estimation of confidence intervals for all contrasts, from the Scheffé viewpoint. Stein's distribution may be used to design experiments for two-stage sampling tests of hypotheses chosen *post hoc* and tested by the R technique. We draw n replicates for each population and use the resulting data to estimate $\sigma_\epsilon{}^2$ by MSW with $\nu_2 = a(n-1)$. We then draw $N-n$ further replicates for each population in which

$$N = [\lambda'^2_{\phi_0}\text{MSW}/\delta'^2] + 1, \tag{33}$$

in which δ' is the value of each alternative to normalized hypotheses (all ${}_1\kappa_i$ are assumed to have the same δ'). Unfortunately, not only has Stein's distribution not been numerically integrated to find values of $\lambda'^2_{\phi_0}$; it does not seem to have been numerically integrated at all; so tables of non-central parameters are not available for practical use, except when $\nu_1 = 1$.

As an example of the application of the R technique consider an experiment designed to investigate the effect of various conditions on size constancy. Six groups will be used to study the effects of reduced and full viewing conditions, visual angle between the standard and the variable and the effects of instructions to set the variable to the "apparent" or the projective size of the standard. The various conditions are shown in Table 2.

TABLE 2. CONDITIONS OF HYPOTHETICAL EXPERIMENT ON SIZE CONSTANCY

	Group					
	1	2	3	4	5	6
Viewing	Reduced	Reduced	Full	Full	Full	Full
Angle	5°	60°	5°	60°	5°	60°
Instructions	Apparent	Apparent	Projective	Projective	Apparent	Apparent

We are interested not only in differences between treatment populations but also in the extent to which the means of populations approach projective equality to the standard (which will happen when the variable is set at 10 cm); so we use a decision set with $\nu_1 = p = 6$. We set $\alpha = 1 - \beta = 0{\cdot}05$ and $\lambda_i^2 = n$. Since we shall use $\hat{\sigma}_\epsilon^2 = \mathrm{MSW}$ in eqn. (10) then $\nu_2 = a(n-1)$ and the various ν_2 in Table 1 represent $n = 1\frac{2}{3} \times \frac{1}{3}$ to $4\frac{1}{3}$, 6, $7\frac{2}{3}$, $9\frac{1}{3}$, 11 and ∞. If we choose $n = 11$, then $\nu_2 = 60$ and $\lambda^2_{0.05^6} = 10{\cdot}754 < n$. Eleven subjects are therefore assigned, at random, to each of the six treatment groups, the experiment is carried out and we observe $\mathrm{MSW} = 25$, $\bar{Y}_j = 11$, 12, 12, 13, 14 and 15 cm. We make $m_j = -10$ cm (to adjust for projective equality) and define $M_j = \bar{Y}_j + m_j = 1$, 2, 2, 3, 4 and 5 cm.

These data yield

$$\mathrm{USSB} = n \sum_j M_j^2 = 649$$

$$\mathrm{SSB} = n \sum_j (M_j - M.)^2 = 119{\cdot}16.$$

When divided by ν_1 $\mathrm{MSW} = 6 \times 25$, these quantities yield 4·33 and 0·79. Since $F_{0\cdot 95^6} = 1{\cdot}31$ we may reject

$$q = [4{\cdot}33/1{\cdot}31] = [3{\cdot}3] = 3$$

null hypotheses. Since $[0{\cdot}79/1{\cdot}31] = 0$, we shall not be able to find any contrasts that can be rejected by the eqn. (10) decision rule. After a few trials of eqn. (10) with the M_j we settle on the following mutually orthogonal hypotheses as our decision set:

$$\begin{array}{llll}
{}_0\kappa_1 = \mu_1 + \mu_2 + \mu_3 = 0, & L_1 = 91{\cdot}\dot{6}, & F_1 = 0{\cdot}61, & \text{Retain } {}_0\kappa_1; \\
{}_0\kappa_2 = \mu_4 + \mu_5 = 0, & L_2 = 269{\cdot}5, & F_2 = 1{\cdot}80, & \text{Reject } {}_0\kappa_2; \\
{}_0\kappa_3 = \mu_6 = 0, & L_3 = 275{\cdot}0, & F_3 = 1{\cdot}83, & \text{Reject } {}_0\kappa_3; \\
{}_0\kappa_4 = \mu_4 - \mu_5 = 0, & L_4 = 5{\cdot}5, & F_4 = 0{\cdot}04, & \text{Retain } {}_0\kappa_4; \\
{}_0\kappa_5 = -2\mu_1 + \mu_2 + \mu_3 = 0, & L_5 = 7{\cdot}\dot{3}, & F_5 = 0{\cdot}05, & \text{Retain } {}_0\kappa_5; \\
{}_0\kappa_6 = \mu_2 - \mu_3 = 0, & L_6 = 0, & F_6 = 0, & \text{Retain } {}_0\kappa_6.
\end{array}$$

When $\lambda^2 = n$ we have $\delta_i = \sigma_\epsilon \sqrt{\sum_j c_{ij}^2}$; so the above decision set gives $\delta_i = 0$, $\sigma_\epsilon \sqrt{2}$, σ_ϵ, 0, 0, 0. When these quantities are inserted in eqn. (17) we obtain $\mu_j = 0$, 0, 0, $\sigma_\epsilon/\sqrt{2}$, $\sigma_\epsilon/\sqrt{2}$, σ_ϵ.

In spite of the fact that we could not *directly* reject $2\mu_1+2\mu_2+2\mu_3-3\mu_4-3\mu_5=0$, this hypothesis is false by implication since its value is $0+0+0-3\sigma_\epsilon/\sqrt{2}-3\sigma_\epsilon/\sqrt{2}=-\sigma_\epsilon\sqrt{18}$. The most plausible reason why this hypothesis could not be directly rejected is that its $\delta=-\sigma_\epsilon\sqrt{18}$ gives $\lambda^2=0{\cdot}6n$ and we were using $\lambda^2=n$: our test did not have enough power to detect directly the fact that this hypothesis is false.

We have not rejected as many hypotheses as the variation among the M_j justifies. This was because we could not find three mutually orthogonal hypotheses which had a simple interpretation and could be rejected. A conservative decision set will commonly occur in the application of the R technique for this reason. There does exist an infinite number of negligibly different sets of six mutually orthogonal hypotheses in each of which there are three hypotheses with $F_i>1{\cdot}31$. One interesting set is that with q values of $F_i=F_m/q$ ($4{\cdot}33/3$ in our example) and the remaining $p-q$ values of $F_i=0$. Such a set is uniquely determined by the rules that it contain mutually orthogonal hypotheses with q of them having $F_i=F_m/q$ and the remainder $F_i=0$. After much tedious algebra it may be shown that this set yields

$$\mu_j=\sigma_\epsilon M_j\sqrt{(q\lambda^2_{(1-\beta)\nu_1}/\text{USSB})}, \tag{34}$$

when its δ_i and ${}_p\mathbf{C}_a$ are inserted in eqn. (17). If the decision set is restricted to contrasts, (34) becomes

$$\mu_j-\mu.=\sigma_\epsilon(M_j-M.)\sqrt{(q\lambda^2{}_{(1-\beta)\nu_1}/\text{SSB})}. \tag{35}$$

In our example, eqn. (34) gives $\mu_j=0{\cdot}226\sigma_\epsilon M_j$.

If all ${}_0\kappa_i$ were true, the probability of rejecting $r=3$ or more of them is the area to the right of $rF_{(1-\alpha)\nu_1}=3\times1{\cdot}31$ in the variance ratio distribution with $\nu_1=p=6$, $\nu_2=a(n-1)=60$ and $\lambda_m{}^2=0$, which is $0{\cdot}002$. If $q=3$ of the alternative hypotheses were true and the remaining ${}_0\kappa_i$ true, the probability of rejecting $r=2$ or fewer ${}_0\kappa_i$ is the area to the left of $(r+1)F_{(1-\alpha)\nu_1}=3\times1{\cdot}31$ in the variance ratio distribution with $\nu_1=p=6$, $\nu_2=a(n-1)=60$ and $\lambda_m{}^2=q\lambda^2_{(1-\beta)\nu_1}=3\times11$, which is $0{\cdot}104$.

There can be no serious objection to using a set of hypotheses which are not mutually orthogonal, but which have ${}_p\mathbf{C}_a$ with rank p, as the formal decision set. To maintain the orthogonal basis for the R method we attempt to reject q, given by eqn. (28), of the ${}_0\kappa_i$ and the non-central parameter for the sub-set of ${}_1\kappa_i$ accepted is $q\lambda^2_{(1-\beta)\nu_1}$. The probabilities of r rejections of ${}_0\kappa_i$ are those given earlier. We will assume that hypotheses are normalized, since this simplifies the problem of finding the δ_i for the ${}_1\kappa_i$ accepted. The δ_i must satisfy

$$q\lambda^2_{(1-\beta)\nu_1}=n_1\boldsymbol{\delta}_p{}^T({}_p\mathbf{C}_a\,{}_a\mathbf{C}_p{}^T)^{-1}{}_p\boldsymbol{\delta}_1\sigma_\epsilon{}^{-2}=n_1\boldsymbol{\delta}_p{}^T{}_p\mathbf{U}_p\,{}_p\boldsymbol{\delta}_1\sigma_\epsilon{}^{-2}. \tag{36}$$

If we accept the earlier view that all $\lambda_i{}^2$ be made equal, this is equivalent to all δ_i being equal in eqn. (36) for ${}_1\kappa_i$ accepted. Of course for ${}_0\kappa_i$ accepted $\delta_i=0$.

With these conditions we find the solution from eqn. (36) to be

$$\delta = \sigma_\epsilon \sqrt{(q\lambda^2_{(1-\beta)\nu_1}/nu)} \tag{37}$$

in which u is the sum of the elements of the sub-matrix of ${}_p\mathbf{U}_p$ which results when the rows and columns of ${}_p\mathbf{U}_p$, corresponding to ${}_0\kappa_i$ accepted, are deleted.

Whereas with mutually orthogonal null hypotheses it is impossible to find more than q (28) which are rejectable by the (10) rule, this *is* possible with null hypotheses which are intercorrelated. We ought to modify the $F_{(1-\alpha)\nu_1}$ criterion for ${}_0\kappa_i$ which are intercorrelated. This could be done by adapting (5), but it seems an unnecessary refinement for practical application.

6. SUMMARY

Given a populations in which the variate Y is distributed normally, with means μ_j $(j=1, 2, \ldots, a)$ and common variance σ_ϵ^2, the decision rule proposed in this paper for linear statistical hypotheses about the μ_j is to reject null hypotheses for which $F_i = L_i/\nu_1\text{MSW} \geqslant F_{(1-\alpha)\nu_1}$. When p null hypotheses, with ${}_p\mathbf{C}_a$ of rank p, are chosen, independently of the data to be used for testing them, $\nu_1 = 1$. If the selection of hypotheses is influenced by the test data, $\nu_1 = p$ and we may reject $q = [F_m/F_{(1-\alpha)\nu_1}] \leqslant \nu_1$ of them.

It is recommended that the sample size (n) be chosen to discriminate null hypotheses from reasonable alternatives with known probability. Tables of non-central parameters which may be used to assist the choice n are available from the author. When one-stage sampling is used, alternatives to null hypotheses must be expressed in σ_ϵ units, but with two-stage sampling these alternatives may be expressed in the measured units of Y. Tables of non-central parameters appropriate to two-stage sampling are available, as yet, only for $\nu_1 = 1$.

The methods given in this paper for testing null hypotheses not only allow us to anticipate the type I error-rate and power (or the expected proportion of rejections of null hypotheses) given statements of what is true in the populations, they also allow us to calculate the probability of r rejections of null hypotheses $(0 \leqslant r \leqslant \nu_1)$.

REFERENCES

FISHER, R. A. (1928). The general sampling distribution of the multiple correlation coefficient. *Proc. Roy. Soc.* A **121**, 654–673.

HEALY, W. C., Jr. (1956). Two-sample procedures in simultaneous estimation. *Ann. math. Statist.* **27**, 687–702.

KENDALL, M. G. and STUART, A. (1961). *The Advanced Theory of Statistics*, Vol. 2. London: Griffin.

MOSHMAN, J. (1958). A method for selecting the size of the initial sample in Stein's two sample procedure. *Ann. math. Statist.* **29**, 1271–1275.

PATNAIK, P. B. (1949). The non-central χ^2 and F-distributions and their applications. *Biometrika* **36**, 202–232.

RODGER, R. S. (1965). *Intermediate Statistics*. Sydney: University Bookshop.

RODGER, R. S. (1967). Type I errors and their decision basis. *Br. J. math. statist. Psychol.* **20,** 51–62.

SCHEFFÉ, H. (1953). A method for judging all contrasts in the analysis of variance. *Biometrika* **40,** 87–104.

SEELBINDER, B. M. (1953). On Stein's two-stage sampling scheme. *Ann. math. Statist.* **24,** 640–649.

STEIN, C. (1945). A two-sample test for a linear hypothesis whose power is independent of the variance. *Ann. math. Statist.* **16,** 243–258.

TANG, P. C. (1938). The power function of the analysis of variance tests. *Statist. res. Mem.* **2,** 126–157.

WISHART, J. (1932). A note on the distribution of the correlation ratio. *Biometrika* **24,** 441–456.

Vol. 20
Part 2
205-215

The British Journal of Mathematical and
Statistical Psychology

November
1967

FACTOR INTERACTION IN NONLINEAR FACTOR ANALYSIS

By Roderick P. McDonald
University of New England, N.S.W., Australia

Factor interaction models are defined as those cases in nonlinear factor analysis in which the specification equation contains products of two or more latent variables or functions of latent variables. A complete algebraic treatment is given for the case of a product of two latent variables, and certain more general cases are briefly outlined.

1. Introduction

McDonald (1962, 1965 a, 1965 b, 1967 a, 1967 b, 1967 c) has developed a treatment of factor analysis which allows for nonlinear, possibly nonmonotonic, relations between the observed variables and one or more latent variables or factors.

Bartlett (1953), in a discussion of the linearity hypothesis in factor analysis, noted that if we include the product of two existing factors as an additional term in the specification equation of linear factor analysis, it will act just like a further common factor as far as the correlation properties of the observed variables are concerned. He further suggested that to discriminate such a product term from a genuine third factor, it might be possible to examine the correlation between the suspect term and the constructed product of its possible component factors. Bartlett noted also, however, that such an inquiry would be complicated by rotational indeterminacy, especially in the presence of other genuine factors.

Let us define a *factor interaction* as a term in the factor model which contains the product of two or more latent variables, or the product of functions of two or more latent variables. The object of this paper is to present an extension of the theory given by McDonald (1962, 1967 b, 1967 c) to deal with certain cases involving factor interaction. An account will be given of the simple case described by Bartlett, followed by an account of the more general theory, and a constructed numerical example of the simple case.

2. The Simple Interaction Model

An orthogonal factor model will be considered in which one of the orthogonal components can be represented as a product of two latent variables. Put

$$z_j = a_{j1}x_1 + a_{j2}x_2 + a_{j3}x_1x_2 + e_j, \qquad j = 1, 2, \ldots, n, \tag{1}$$

where a_{j1}, a_{j2}, a_{j3} and $\sigma_j^2 = E(e_j^2)$ are the model parameters, and z_j, x_1, x_2, e_j are random variables. Further, each z_j is a standardized observable variable, i.e., $E(z_j) = 0$, $E(z_j^2) = 1$, $j = 1, 2, \ldots, n$; x_1, x_2 are standardized mutually statistically independent latent variables; each e_j is a random deviate, i.e., $E(e_j)=0$; and the e_j, $j=1, 2, \ldots, n$, are statistically independent of each other and of x_1, x_2.

Under the strong assumption that the latent variables x_1, x_2 are mutually statistically independent, the three components x_1, x_2 and x_1x_2 are indeed orthogonal, since

$$E(x_1 \cdot x_1x_2)=E(x_1^2)E(x_2)=0$$

and

$$E(x_2 \cdot x_1x_2)=E(x_1)E(x_2^2)=0.$$

Also, the component x_1x_2 is normalized, since

$$E((x_1x_2)^2)=E(x_1^2)E(x_2^2)=1.$$

Model (1) may be contrasted with an alternative

$$z_j=a_{j1}x_1+a_{j2}x_2+a_{j3}x_3+e_j, \tag{2}$$

with analogous notation, where each pair of the latent variables x_1, x_2, x_3 is statistically independent.

Both model (1) and model (2) are consistent with the weaker model of orthogonal factor analysis

$$z_j=b_{j1}v_1+b_{j2}v_2+b_{j3}v_3+e_j, \tag{3}$$

where it is specified only that

$$\begin{aligned} E(v_pv_q)&=1, \text{ for all } p=q \\ &=0, \text{ for all } p\neq q \end{aligned}$$

and

$E(v_pe_j)=0$, for all p, j, but v_1, v_2, v_3 are not necessarily statistically independent of each other.

Two $n\times 3$ matrices of *factor loadings* **A** and **B** are now defined whose rows are given respectively by $[a_{j1}, a_{j2}, a_{j3}]$ and $[b_{j1}, b_{j2}, b_{j3}]$. These matrices are assumed to be of full column rank. There exists a 3×3 orthogonal matrix **L** such that

$$\mathbf{A}=\mathbf{BL}, \tag{4}$$

and, correspondingly, for model (1)

$$[x_1, x_2, x_1x_2]=[v_1, v_2, v_3]\mathbf{L}', \tag{5}$$

while for model (2)

$$[x_1, x_2, x_3]=[v_1, v_2, v_3]\mathbf{L}'. \tag{6}$$

That is, supposing that model (3) applies to a given population, then it may also be the case that one of the stronger models (1) or (2) applies to that population. In either of these cases, the relation between the factor loadings is given by eqn. (4). If model (1) applies, the matrix **L** is determined uniquely except for trivial permutations or reflections by certain joint moments of the orthogonal components $[v_1, v_2, v_3]$, of higher order than the second. Under the additional assumption that the third moments of x_1 and x_2 vanish, simple relations can be obtained yielding an analytic solution for rotations in the plane.

If, in particular, the orthogonal transformation **L** corresponds to a single rotation of $[v_1, v_2]$ through an angle of θ_{12}, it is easily shown that

$$\cos 2\theta_{12} = \mathrm{E}(v_1 v_2 v_3) \tag{7}$$

and

$$\sin 2\theta_{12} = \mathrm{E}(v_2{}^2 v_3) = -\mathrm{E}(v_1{}^2 v_3).$$

Thus, from

$$[v_1, v_2, v_3] = [x_1, x_2, x_1 x_2] \begin{bmatrix} c_{12} & s_{12} & 0 \\ -s_{12} & c_{12} & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{8}$$

where $c_{12} = \cos \theta_{12}$, $s_{12} = \sin \theta_{12}$, it follows that

$$\begin{aligned} \mathrm{E}(v_1 v_2 v_3) &= c_{12}{}^2\, \mathrm{E}(x_1{}^2 x_2{}^2) - s_{12}{}^2\, \mathrm{E}(x_1{}^2 x_2{}^2) \\ &= \cos 2\theta_{12}. \end{aligned} \tag{9}$$

Expressions analogous to eqn. (7) can be derived for planar rotations of $[v_1, v_3]$ and $[v_{2,}\ v_3]$, viz.,

$$\begin{aligned} \mathrm{E}(v_1 v_2 v_3) &= \cos 2\theta_{13} - \cos \theta_{13} \sin \theta_{13}\, \mathrm{E}(x_2{}^3) \\ &= \cos 2\theta_{13}, \end{aligned} \tag{10}$$

under the assumption that

$$\mathrm{E}(x_2{}^3) = 0,$$

and

$$\begin{aligned} \mathrm{E}(v_1 v_2 v_3) &= \cos 2\theta_{23} - \cos \theta_{23} \sin \theta_{23}\, \mathrm{E}(x_1{}^3) \\ &= \cos 2\theta_{23}, \end{aligned} \tag{11}$$

under the assumption that

$$\mathrm{E}(x_1{}^3) = 0.$$

Alternatively, a least-squares procedure can be developed which corresponds to the theory given by McDonald (1962).

Writing

$$[v_1{}^*, v_2{}^*, v_3{}^*] = [c_{12} v_1 - s_{12} v_2,\ s_{12} v_1 + c_{12} v_2,\ v_3], \tag{12}$$

θ_{12} is chosen to minimize the function

$$\begin{aligned} \Phi &= \mathrm{E}((v_3{}^* - v_1{}^* v_2{}^*)^2) \\ &= \mathrm{E}(\{v_3 - (c_{12} v_1 - s_{12} v_2)(s_{12} v_1 + c_{12} v_2)\}^2). \end{aligned} \tag{13}$$

It may be noted that if the model holds precisely in a given population, the expression inside the braces is identically zero for all x_1, x_2, if a rotation is chosen such that $v_1{}^* = x_1$, $v_2{}^* = x_2$. Hence, the minimum value of Φ will be precisely zero in such a population. Further,

$$\mathrm{E}(v_3{}^{*2}) = \mathrm{E}(v_1{}^{*2}\, v_2{}^{*2}) = 1, \tag{14}$$

whence

$$\begin{aligned} \Phi &= 2 - 2\mathrm{E}(v_1{}^* v_2{}^* v_3{}^*) \\ &= 2 - 2\mathrm{E}(c_{12} s_{12}(v_1{}^2 v_3 - v_2{}^2 v_3) + (c_{12}{}^2 - s_{12}{}^2) v_1 v_2 v_3), \end{aligned} \tag{15}$$

so that the function Φ will be a minimum when the trivariate moment $E(v_1{}^*v_2{}^*v_3{}^*)$ is a maximum.

Setting the partial derivative of Φ with respect to θ_{12} equal to zero in eqn. (15), it is found that

$$\tan 2\theta_{12} = \frac{E(v_2{}^2v_3) - E(v_1{}^2v_3)}{2E(v_1v_2v_3)} \tag{16}$$

and by symmetry, for single rotations of $[v_1, v_3]$, $[v_2, v_3]$ through angles θ_{13}, θ_{23} respectively,

$$\tan 2\theta_{13} = \frac{E(v_3{}^2v_2) - E(v_1{}^2v_2)}{2E(v_1v_2v_3)} \tag{17}$$

and

$$\tan 2\theta_{23} = \frac{E(v_3{}^2v_1) - E(v_2{}^2v_1)}{2E(v_1v_2v_3)}. \tag{18}$$

Eqns. (16), (17) and (18) may be compared with eqns. (7), (10) and (11). The former equations, together with the conditions for a minimum, determine the angles uniquely within a range of 180°. This restriction of range corresponds to the trivial indeterminacy of **L** noted above.

By a rotation of x_1 and x_2 through an angle of 45°, model (1) can be rewritten in the form

$$z_j = a_{j1}\frac{(x_1{}^* - x_2{}^*)}{\sqrt{2}} + a_{j2}\frac{(x_1{}^* + x_2{}^*)}{\sqrt{2}} + a_{j3}\frac{(x_1{}^{*2} - x_2{}^{*2})}{\sqrt{2}} + e_j. \tag{19}$$

By rearranging terms in this equation, the model may then be expressed as the sum of a quadratic function in $x_1{}^*$ and a quadratic function in $x_2{}^*$, so that it might appear to be transformable into a member of the class of polynomial models treated in the author's earlier papers. However, it is a degenerate case of this class, since the coefficients in the rearranged equation are not linearly independent. (The coefficients of $x_1{}^{*2}$ and $x_2{}^{*2}$ differ only in sign.) Eqn. (19) serves to show that the possession of a product term x_1x_2 is not an invariant property of the model, so that an interpretation in terms of the action of one of the latent variables as a *moderator variable* (cf. Saunders, 1956) is not entailed by its mathematical properties. However, the degenerate polynomial case represented by eqn. (19) would certainly appear less parsimonious than the original form of the model.

3. More General Cases

The following model is now considered:

$$z_j = a_{j1}x_1 + a_{j2}x_2 + a_{j3}x_1x_2 + a_{j4}x_3 + e_j, \tag{20}$$

with the previous notation, where a fourth orthogonal component has been added to the terms of model (1), viz., a latent variable x_3 which is assumed statistically independent of x_1 and of x_2.

Again comparing the model (20) with a weaker model

$$z_j = b_{j1}v_1 + b_{j2}v_2 + b_{j3}v_3 + b_{j4}v_4 + e_j, \tag{21}$$

whose components v_1, v_2, v_3, v_4 are mutually orthogonal, it is evident that the arguments of Section 2 will hold for simple rotations of v_1 with v_2, v_1 with v_3, and v_2 with v_3. In addition, it is easy to develop a least-squares procedure analogous to the above, for simple rotations of v_1 with v_4, v_2 with v_4, and v_3 with v_4. Thus for the case of a single rotation of $[v_1, v_4]$, it is possible to write

$$\begin{aligned}\Phi &= 2 - 2\mathrm{E}(v_1{}^*v_2{}^*v_3{}^*)\\ &= 2 - 2\mathrm{E}((c_{14}v_1 - s_{14}v_4)v_2v_3)\\ &= 2 - 2c_{14}\mathrm{E}(v_1v_2v_3) + 2s_{14}\mathrm{E}(v_2v_3v_4).\end{aligned} \tag{22}$$

Setting the partial derivative of Φ with respect to θ_{14} equal to zero yields

$$\tan\theta_{14} = -\frac{\mathrm{E}(v_2v_3v_4)}{\mathrm{E}(v_1v_2v_3)} \tag{23}$$

which, together with the condition for a minimum, determines θ_{14} uniquely.

By symmetry,

$$\tan\theta_{24} = -\frac{\mathrm{E}(v_1v_3v_4)}{\mathrm{E}(v_1v_2v_3)}, \tag{24}$$

$$\tan\theta_{34} = -\frac{\mathrm{E}(v_1v_2v_4)}{\mathrm{E}(v_1v_2v_3)}. \tag{25}$$

As noted earlier, the transformation obtained by maximizing the prescribed trivariate moment is determined only up to an arbitrary permutation of the elements of the model as given by the expression on the left in eqn. (5), say. Suppose that, given such a transformation on a finite sample of factor scores, it is desired to determine which column represents the product term. In the absence of ' errors ' in the component variates due to unique variations (a situation which does not obtain in practice; see McDonald, 1962) this decision is easily made by inspection of the transformed matrix in graphical or tabular form. Also, in the absence of such ' errors ', a number of asymmetries in the magnitudes of the joint or univariate moments could be employed as a numerical criterion for identification of the product term. For example, it is immediately obvious that the fourth moment of the product term equals the product of the fourth moments of its arguments. However, the effects on such quantities of ' errors ' due to unique variance are quite complex, and no strictly unequivocal decision criterion has yet been found.

The relations supplied by eqns. (16), (17), (18), (23), (24) and (25) have been combined in a computer program FAINT (McDonald, 1965 b, 1967 a) to form an iterative algorithm which seeks to rotate a sample matrix of *factor scores*, so that the sample trivariate moment within some group of three successive columns of N observations on $\mathbf{v}$,

$$m_{111} = \frac{1}{N}\left(\sum_{i=1}^{N} v_{i,k}v_{i,k+1}v_{i,k+2}\right) \tag{26}$$

will be a maximum or, what is the same thing, the sample analogue of Φ as given by eqn. (13) will be a minimum. In practice, it is recommended that sample *component variates* as defined by McDonald (1962) be employed. It is easily shown that the relevant sample moments of these are unbiased estimates of the corresponding moments of the 'true' factor scores. If the data conform to a model such as

$$z_j = a_{j1}x_1 + a_{j2}x_2 + a_{j3}x_1x_2 + a_{j4}x_3 + a_{j5}x_4 + a_{j6}x_3x_4 + \ldots e_j, \tag{27}$$

a rotation will be applied to the estimated factor score matrix which enables the prescribed relations between the columns of the matrix to be detected. The rotation obtained by maximizing m_{111} can then be applied to the factor loadings to complete the estimation of the model parameters.

The program, at any iteration, uses relations of the type of eqns. (16), (17) and (18), respectively, to rotate columns k with $k+1$, k with $k+2$, and $k+1$ with $k+2$, and uses relations of the type of eqns. (23), (24) and (25) to rotate each of the columns k, $k+1$ and $k+2$ with each other column of the factor score matrix. This procedure is iterated until no rotation of pairs is indicated to be necessary, of greater than a size prescribed in degrees by the user.

The above remarks generalize readily to the model

$$z_j = a_{j1}x_1 \ldots + a_{j,k}x_k + a_{j,k+1}x_{k+1} + a_{j,k+2}x_{k+2} + a_{j,k+3}x_kx_{k+1}x_{k+2} + \ldots e_j, \tag{28}$$

and the like, in which triple products or higher order products of latent variables are present as orthogonal components. An important restriction on these models is that each of the factors making up the product must also be present as a separate orthogonal component of the model. Typically, if the model contains an orthogonal component which consists of a product of m terms, we would seek a rotation such that $m+1$ successive columns of the sample factor score matrix yield a maximum sample moment on $(m+1)$ variates, of the type

$$m_{11\ldots1} = \mathrm{E}(v_kv_{k+1}\ldots v_{k+m}). \tag{29}$$

Finally models of the following type are considered:

$$z_j = \ldots a_{jp}h_p(x_1) + a_{jr}h_q(x_2) + a_{jt}h_p(x_1)h_q(x_2) + \ldots e_j, \tag{30}$$

where

$$\begin{aligned}\mathrm{E}(h_p(x_1)h_q(x_1)) &= 1,\ p = q\\ &= 0,\ p \neq q.\end{aligned}$$

That is, in place of the latent variables x_1, x_2, . . ., there are a number of *orthonormal functions* in each latent variable. A treatment of nonlinear factor analysis, without interaction terms, for the case of orthonormal polynomials, has been given by McDonald (1962, 1965 a, 1967 c). For the present case, it can be noted immediately that the methods applied to model (27) will apply directly to model (30), provided that each polynomial in the product term or terms is also present in the model as a single orthogonal component. A general computer program for nonlinear factor analysis, which includes optional specification of models of the type of eqns. (28) and (30), without any restrictions as above on the univariate moments of the latent variables, is currently under development.

4. A CONSTRUCTED NUMERICAL EXAMPLE

In this section a numerical example corresponding to model (20) is given. For this and other applications, a general test program, ARTY, has been written to generate a matrix **H** of factor scores, which will have the properties prescribed by a specific nonlinear factor model. The number of columns (factors) r and the number of rows (cases) $n > r$ are prescribed, and the elements of each column of the matrix are generated, either as sample values of a normally distributed random variable, or as values of a constructed polynomial of prescribed degree in a column previously generated orthonormal in the sample, or as a product of two or more previously generated columns. The program then generates a random orthonormal $r \times r$ matrix **L**, and finally postmultiplies **H** by **L**, yielding an $n \times r$ matrix **V** of randomly rotated factor scores. These are punched on cards in a format suitable as input to the nonlinear factor analysis programs (McDonald, 1965 b, 1967 a).

A 50×4 matrix, generated by program ARTY, is shown in Table 1. This matrix was constructed as a random rotation of a matrix corresponding to the model (x_1, x_2, x_1x_2, x_3). A fifth column on the right of the matrix contains the constructed product of corresponding elements in the first and second columns of the matrix. The correlation between this constructed product and column three of the matrix is $-0{\cdot}284$.

TABLE 1. INPUT FACTOR SCORES AND CONSTRUCTED PRODUCT OF FIRST AND SECOND FACTOR SCORE

	I	II	III	IV	I × II
1	−4·25	2·98	1·01	3·77	−12·69
2	−0·31	1·70	−1·47	1·22	−0·55
3	0·54	0·74	−0·26	−0·89	0·41
4	0·15	−0·60	0·67	0·52	−0·09
5	0·00	−0·05	1·06	−0·29	0·00
6	0·63	0·85	−0·61	−0·76	0·55
7	0·18	0·40	0·65	−0·85	0·08
8	0·67	0·92	−0·82	−0·65	0·63
9	−0·79	−1·74	−2·13	0·07	1·38
10	0·57	1·22	−1·44	0·01	0·70
11	−0·78	−1·75	−2·13	0·08	1·37
12	0·00	−0·14	1·05	−0·16	0·00
13	0·30	0·52	0·38	−0·93	0·16
14	0·53	−1·40	−0·87	1·34	−0·74

(Continued)

TABLE 1 (*continued*)

	I	II	III	IV	I×II
15	0·15	−0·62	0·66	0·54	−0·10
16	0·26	−0·82	0·36	0·81	−0·22
17	0·15	1·49	−1·58	0·70	0·22
18	0·03	−0·32	0·96	0·11	−0·01
19	0·34	0·57	0·26	−0·94	0·20
20	0·30	0·52	0·38	−0·93	0·16
21	−1·20	−1·70	−2·08	−0·31	2·06
22	0·00	−0·19	1·04	−0·08	0·00
23	0·00	−0·01	1·06	−0·35	0·00
24	0·00	−0·02	1·06	−0·33	0·00
25	0·20	−0·71	0·52	0·67	−0·15
26	0·04	−0·33	0·96	0·12	−0·01
27	0·09	−0·48	0·82	0·34	−0·04
28	0·66	0·89	−0·74	−0·70	0·59
29	0·01	0·09	1·02	−0·51	0·00
30	0·51	−1·54	−1·24	1·34	−0·81
31	0·07	0·24	0·89	−0·70	0·02
32	0·63	1·15	−1·35	−0·15	0·74
33	0·44	1·33	−1·54	0·28	0·59
34	0·56	0·77	−0·34	−0·86	0·44
35	0·33	−0·94	0·14	0·95	−0·31
36	0·07	−0·43	0·87	0·27	−0·03
37	0·38	−1·03	−0·03	1·06	−0·41
38	0·05	−0·36	0·93	0·17	−0·01
39	0·03	−0·30	0·98	0·08	−0·01
40	0·29	0·51	0·40	−0·92	0·15
41	0·37	−1·01	0·00	1·03	−0·38
42	0·14	0·34	0·75	−0·81	0·07
43	0·60	1·19	−1·41	−0·05	0·72
44	0·26	0·48	0·48	−0·91	0·13
45	0·19	−0·69	0·56	0·63	0·13
46	0·66	0·90	−0·77	−0·68	0·61
47	0·41	−1·09	−0·14	1·11	−0·45
48	0·18	−0·66	0·59	0·60	−0·12
49	0·03	0·15	0·98	−0·58	0·01
50	−4·83	−1·04	−0·62	−3·48	5·03

The 50×4 matrix as shown in Table 1 was input to program FAINT, with a prescribed precision of two degrees. That is, the criterion for convergence was that no rotation of one column with another through an angle greater than two degrees be required. The rotated factor score matrix, again with a fifth column on the right containing the constructed product of elements from the first and second columns, is given in Table 2. The correlation between this constructed product and column three of the matrix is now 0·987. The required rotation is given in Table 3.

TABLE 2. ROTATED FACTOR SCORES AND CONSTRUCTED PRODUCT OF FIRST AND SECOND FACTOR SCORES

	I	II	III	IV	I × II
1	−3·23	3·14	−4·67	0·32	−10·14
2	−0·23	1·94	−0·46	1·63	−0·45
3	1·09	0·57	0·40	−0·15	0·62
4	−0·84	−0·43	0·21	−0·42	0·36
5	−0·10	−0·15	−0·07	−1·08	0·02
6	1·12	0·73	0·51	0·19	0·82
7	0·66	0·16	0·05	−0·94	0·11
8	1·11	0·84	0·55	0·42	0·93
9	0·20	−1·70	−0·33	2·27	−0·34
10	0·77	1·34	0·46	1·19	1·03
11	0·19	−1·70	−0·32	2·28	−0·32
12	−0·23	−0·20	−0·05	−1·03	0·05
13	0·85	0·29	0·16	−0·73	0·25
14	−1·24	−0·88	0·84	1·33	1·09
15	−0·85	−0·44	0·33	−0·40	0·38
16	−1·04	−0·54	0·39	−0·02	0·56
17	0·24	1·67	0·03	1·54	0·40
18	−0·50	−0·30	0·02	−0·84	0·15
19	0·91	0·34	0·21	−0·63	0·31
20	0·85	0·29	0·16	−0·73	0·25
21	0·54	−1·83	−0·76	2·14	−0·99
22	−0·31	−0·23	−0·03	−0·98	0·07
23	−0·03	−0·13	−0·08	−1·10	0·00
24	−0·05	−0·14	−0·07	−1·10	0·00
25	−0·95	−0·49	0·30	−0·22	0·47
26	−0·51	−0·30	0·02	−0·83	0·15
27	−0·69	−0·37	0·12	−0·63	0·26
28	1·12	0·79	0·53	0·32	0·88
29	0·14	−0·06	−0·08	−1·13	−0·01
30	−1·15	−1·00	0·88	1·70	1·15
31	0·40	0·03	−0·04	−1·09	0·01
32	0·88	1·19	0·52	1·04	1·05
33	0·58	1·45	0·32	1·36	0·84
34	1·10	0·60	0·43	−0·08	0·66
35	−1·13	−0·60	0·49	0·22	0·68
36	−0·63	−0·35	0·08	−0·70	0·22
37	−1·18	−0·65	0·57	0·43	0·77
38	−0·55	−0·32	0·04	−0·79	0·18
39	−0·47	−0·29	0·01	−0·86	0·14
40	0·84	0·28	0·15	−0·75	0·24
41	−1·17	−0·63	0·55	0·38	0·74
42	0·57	0·12	0·01	−1·01	0·07
43	0·82	1·25	0·49	1·13	1·03
44	0·79	0·24	0·12	−0·81	0·19

(Continued)

TABLE 2 (*continued*)

	I	II	III	IV	I × II
45	−0·92	−0·47	0·27	−0·28	0·43
46	1·12	0·81	0·54	0·36	0·91
47	−1·21	−0·68	0·62	0·55	0·82
48	−0·90	−0·46	0·26	−0·32	0·42
49	0·24	−0·02	−0·07	−1·13	0·00
50	2·96	−2·62	−4·61	0·03	−7·76

TABLE 3. ROTATION

0·035	0·983	0·153	−0·091
−0·334	−0·071	−0·029	−0·940
0·234	−0·116	0·952	−0·100
−0·912	0·021	0·261	0·314

In a real case, the factor score matrix of Table 1 would have been obtained as the end result of an orthogonal factor analysis of m observed variables. To complete the estimation of the model parameters, the $m \times 4$ matrix of factor loadings obtained in the preliminary factor analysis would be operated upon, with the orthonormal matrix given in Table 3. Clearly, the interpretation is that the first, second and fourth orthogonal components are mutually statistically independent latent variables, whereas the third represents a simple interaction of the first two.

5. GENERAL DISCUSSION

A good deal of experience with empirical data will be necessary before it is possible to say whether the models described above will commonly prove to be necessary and useful to the psychologist.

There is some danger that the simple interaction model may have spurious appeal for those psychologists who tend to reify factors as something akin to faculties, because it seems to describe interactions of organismic functions. The fact that model (1) can be transformed into the form given in eqn. (19), in which there is no product term, deserves some further emphasis. The alternative interpretation of data in terms of model (19) would entail a radically different psychological theory, and a choice between model (1) and model (19) may rest on psychological considerations, if, indeed, a rational choice can be made.

ACKNOWLEDGEMENTS

This work was done while the author was a Visiting Research Fellow at the Educational Testing Service. It was supported in part by Office of Naval Research Contract Nonr 1858-(15), Project Designation NR 150-088, and National Science Foundation Grant, GB-3402, continuation of G-22889, Project on " Mathematical Techniques in Psychology ", Principal Investigator, Professor Harold Gulliksen, Psychology Department, Princeton University.

The work made use of computer facilities at Princeton University that are supported in part by National Science Foundation Grant NSF-GP 579. It is based in part on a paper read to the Psychometric Society at Niagara Falls, October 1964.

REFERENCES

BARTLETT, M. S. (1953). Factor analysis in psychology as a statistician sees it. In: Uppsala symposium on psychological factor analysis. *Nord. Psychol. Monog.* No. 3.

MCDONALD, R. P. (1962). A general approach to nonlinear factor analysis. *Psychometrika* **27,** 397–415.

MCDONALD, R. P. (1965 a). Difficulty factors and nonlinear factor analysis. *Br. J. math. statist. Psychol.* **18,** 11–23.

MCDONALD, R. P. (1965 b). Computer programs for nonlinear factor analysis. *Research Memorandum* 65–11. Princeton: Educational Testing Service.

MCDONALD, R. P. (1967 a). Some IBM 7090–94 programs for nonlinear factor analysis. *Behavl. Sci.* **12,** 72–74.

MCDONALD, R. P. (1967 b). Nonlinear factor analysis. *Psychometric Monogr.* (in press).

MCDONALD, R. P. (1967 c). Numerical methods for polynomial models in nonlinear factor analysis. *Psychometrika* **32,** 77–112.

SAUNDERS, D. R. (1956). Moderator variables in prediction. *Educ. psychol. Measur.* **16,** 209–222.

Vol. 20
Part 2
217–225

The British Journal of Mathematical and
Statistical Psychology

November
1967

A PROCEDURE FOR DETERMINING SLATER'S i AND ALL NEAREST ADJOINING ORDERS

By J. P. N. Phillips
University of Hull

A new procedure is described for determining Slater's i and all nearest adjoining orders. The validity of the procedure is proved, and it is compared with Remage and Thompson's algorithm for the same purpose.

1. Introduction

The classical measure of internal consistency for a complete table of pair-comparison preference judgements is the number of circular triads, Kendall's d; and the classical method of assigning ranks to the objects in an inconsistent table is to take those (sometimes tied) given by the row totals. Recently, however, Slater (1960, 1961) has suggested as a measure of internal consistency the statistic i, the size of the smallest set or sets of changes in the table which render it consistent, and as rankings for an inconsistent table the nearest adjoining orders, given by actually carrying out these sets of changes.

Unfortunately the practical determination of i and the nearest adjoining orders is a matter of some difficulty. Slater (1961) quotes two rules due to Alway, but Remage and Thompson (1966) have shown, by a counter-example, that they are not valid. The latter authors describe an elegant algorithm, whose validity they prove, but it is somewhat laborious to execute. This note presents a different procedure which has been incorporated in two computer programs reported elsewhere in this *Journal* (Phillips, 1967).

2. The Procedure

Since the example of Kendall (1948) used by Remage and Thompson to demonstrate their algorithm is too simple to display fully the present procedure, a more complicated example will be used as an illustration (see Table 1 (a)).

Step 1 Scan the pair-comparison table and list all circular triads (Table 1 (b)). To render the table consistent, it is necessary to break each triad by changing precisely one or two of the cells occurring in it.

Step 2 List, for each cell in the pair-comparison table containing a plus sign, the number of circular triads in which it occurs and, for convenience, rearrange the list in decreasing order of these frequencies (Table 1 (c)): it can be shown

TABLE 1.

(*a*)

Pair comparisons

	1	2	3	4	5	6	7
1	.	+	+	−	+	−	−
2	−	.	−	+	−	+	+
3	−	+	.	−	−	+	+
4	+	−	+	.	−	+	+
5	−	+	+	+	.	−	+
6	+	−	−	−	+	.	−
7	+	−	−	−	−	+	.

(*b*)

Triads

1	2	4
1	2	6
1	2	7
1	3	6
1	3	7
1	5	4
1	5	7
2	4	3
2	6	5
3	6	5
4	6	5
5	7	6

(*c*) *Frequency of triads per cell* *Cell*	*f*	(*d*) *Cumulative frequency of triads*
6, 5	4	4
1, 2	3	7
7, 1	3	10
1, 3	2	12
1, 5	2	
2, 4	2	
2, 6	2	
3, 6	2	
4, 1	2	
5, 4	2	
5, 7	2	
6, 1	2	
2, 7	1	
3, 2	1	
3, 7	1	
4, 3	1	
4, 6	1	
5, 2	1	
5, 3	1	
7, 6	1	

that, both here and subsequently, cells not in triads may be ignored; the proof, although straightforward, is somewhat messy, and is accordingly omitted. We seek all smallest subsets of cells such that just one or two members of the subset occur in each triad.

Step 3 Determine these smallest subsets by systematic trial and error. As a preliminary, take cumulative totals of the ordered cell frequencies downwards until a total is reached which is equal to or greater than the number of circular triads (Table 1 (*d*)); the number of cells making up this total (here 4) gives a minimum for i.

Then, for all trial values of i from this minimum upwards, until i has been determined, enter the first branch of the following:

Tree-search For each *feasible* (i.e. not clearly impossible, as determined by appropriate cumulative totals of the current ranked cell frequency list) combination of cells of the highest frequency, set up a new list of triads which would remain after all cells of the combination were changed. (The possibility that the changes may generate new circular triads is ignored at this stage.) For

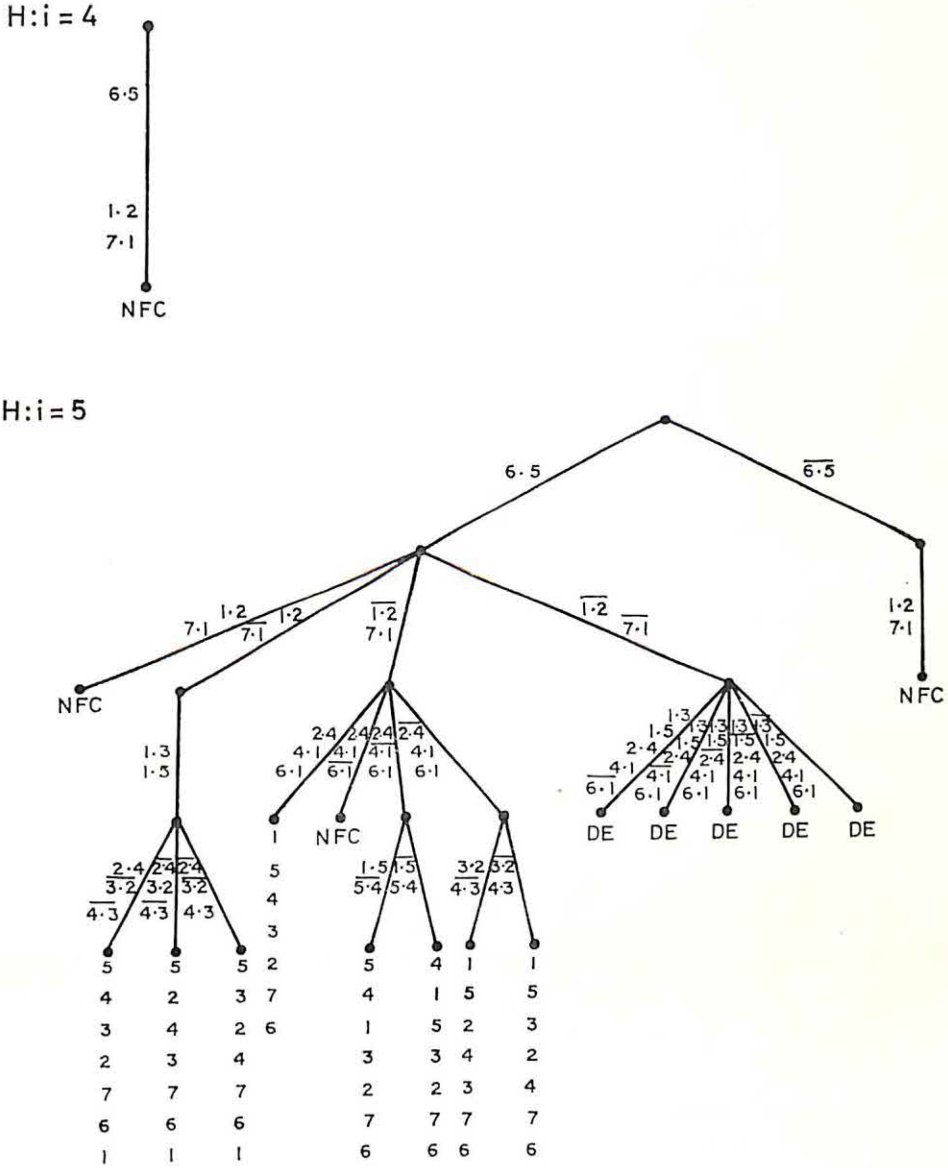


FIGURE 1. Illustration of the tree-search. The two trees illustrate the search procedure (for details see text) on the data of Table 1 under the hypotheses that $i = 4$ (falsified) and that $i = 5$ (verified). Each branch represents an investigation of the effect, on the current list of circular triads, of changing the cells written beside it (e.g., 1,2 $\overline{7,1}$ means " try changing cell 1,2 and not cell 7,1 "), and terminates either in a choice point leading to other branches, or in a dead end (DE), or no feasible combinations (NFC) or a nearest adjoining order.

each such list there are three possibilities. First, it may be empty, in which case verify, by carrying out all changes under consideration, that they have not generated any new circular triads or changed all the cells of a circular triad (practical experience with both randomly generated and experimentally obtained data suggests that this undesired complication rarely occurs); if they have not, note that *i* has been determined (if this has not already been done) and write down the nearest adjoining order obtained; if they have, do nothing. Second, the new list of circular triads may be non-empty, and there may be one or more further changes still available, in which case set up a new ranked cell frequency list for the list of triads remaining (excluding all cells previously of the highest frequency) and enter a new branch of the tree-search with these new data. Third, the new list of circular triads may not be empty, and there may be no more changes available, in which case a dead end has been reached, so do nothing. When all the combinations have been considered, and appropriate action taken in each case, exit from the current branch of the tree-search.

This procedure is illustrated in Figure 1 and Tables 2 to 7.

TABLE 2

(*a*)

Triads		
1	2	4
1	2	6
1	2	7
1	3	6
1	3	7
1	5	4
1	5	7
2	4	3

(*b*)

Cell	*f*
1, 2	3
7, 1	3
1, 3	2
1, 5	2
2, 4	2
4, 1	2
6, 1	2
2, 6	1
3, 6	1
5, 4	1
5, 7	1
2, 7	1
3, 2	1
3, 7	1
4, 3	1

Giving *i* the initial value 4, it can be seen from Table 1 (*c*) that this could be attained only if cell 6,5 were changed. The list of triads then remaining is shown in Table 2 (*a*): since this is non-empty, and there are three further changes still available, a new ranked cell frequency list is set up (Table 2 (*b*)). Re-entering the search with the data of Table 2, it is seen that the eight triads remaining could be broken with the three changes still available only if cells 1,2 and 7,1 were changed. The results of doing so are shown in Table 3, from which it is seen that there is no feasible cell of the highest frequency, so it is necessary to exit from this branch, and then from all higher branches of the search.

TABLE 3

(*a*)

Triads		
1	3	6
1	5	4
2	4	3

(*b*)

Cell	*f*
1, 3	1
1, 5	1
2, 4	1
4, 1	1
6, 1	1
3, 6	1
5, 4	1
3, 2	1
4, 3	1

Then giving *i* the trial value 5, the first branch of the tree-search is re-entered. The successive stages will now be presented in abbreviated form. There are first of all two feasible combinations, changing and not changing 6,5: each will be considered in turn.

Stage 1 Change 6,5 (four changes still available). Result: Table 2; and there are four feasible combinations, changing both, just one, or neither of cells 1,2 and 7,1; consider each in turn.

Stage 1.1 Change 6,5: 1,2; 7,1 (two changes still available). Result: Table 3; and no feasible combinations.

Stage 1.2 Change 6,5: 1,2; not 7,1 (three changes still available). Result: Table 4; and one feasible combination, cells 1,3 and 1,5.

TABLE 4

(*a*)

Triads		
1	3	6
1	3	7
1	5	4
1	5	7
2	4	3

(*b*)

Cell	*f*
1, 3	2
1, 5	2
2, 4	1
4, 1	1
6, 1	1
3, 6	1
5, 4	1
5, 7	1
3, 2	1
3, 7	1
4, 3	1

Stage 1.2.1 Change 6,5: 1,2; not 7,1: 1,3; 1,5 (one change still available). This leaves the single triad 2,4,3 and the three cells which occur in it, each with frequency 1. Changing 2,4 breaks 2,4,3, so we try the whole set of changes 6,5; 1,2; 1,3; 1,5; 2,4: it is found to give the nearest adjoining order 5,4,3,2,7,6,1, so note that $i=5$. Similarly, changing 3,2 gives the order 5,2,4,3,7,6,1 and changing 4,3 gives 5,3,2,4,7,6,1.

Stage 1.3 Change 6,5: not 1,2; 7,1 (three changes still available). Result: Table 5; and four feasible combinations, all three or just two, of cells 2,4, 4,1 and 6,1: consider each in turn.

TABLE 5

(*a*)

Triads		
1	2	4
1	2	6
1	3	6
1	5	4
2	4	3

(*b*)

Cell	*f*
2 4	2
4, 1	2
6, 1	2
1, 3	1
1, 5	1
2, 6	1
3, 6	1
5, 4	1
3, 2	1
4, 3	1

Stage 1.3.1 Change 6,5: not 1,2; 7,1: 2,4; 4,1; 6,1 (no more changes available). This procedure breaks all remaining triads, so it is necessary to try the whole set of changes. It gives the nearest adjoining order 1,5,4,3,2,7,6.

Stage 1.3.2 Change 6,5: not 1,2; 7,1: 2,4; 4,1; not 6,1 (one change still available). Result: Table 6 and no feasible combinations.

TABLE 6

(*a*)

Triads		
1	2	6
1	3	6

(*b*)

Cell	*f*
1, 3	1
2, 6	1
3, 6	1

Stage 1.3.3 Change 6,5: not 1,2; 7,1: 2,4; not 4,1; 6,1 (one change still available). This leaves the single triad 1,5,4 and the cells 1,5 and 5,4, each with frequency 1. Changing each of them in turn gives the nearest adjoining orders 5,4,1,3,2,7,6 and 4,1,5,3,2,7,6.

Stage 1.3.4 Change 6,5: not 1,2; 7,1: not 2,4; 4,1; 6,1 (one change still available). This leaves the single triad 2,4,3 and the cells 3,2 and 4,3, each with frequency 1. Changing each of them in turn gives the nearest adjoining orders 1,5,2,4,3,7,6 and 1,5,3,2,4,7,6.

Stage 1.4 Change 6,5: not 1,2; not 7,1 (four changes still available). Result: Table 2 (*a*) and Table 2 (*b*) with the first two cells missing. Five feasible combinations, changing just four of the cells of frequency 2. Each leads to a dead end.

Stage 2 Do not change 6,5 (five changes still available). Result: Table 1 (*b*) and Table 1 (*c*) with the first cell missing. One feasible combination changing cells 1,2 and 7,1.

Stage 2.1 Do not change 6,5: but change 1,2; 7,1 (three changes still available). Result: Table 7 and no feasible combinations.

TABLE 7

(*a*)

Triads		
1	3	6
1	5	4
2	4	3
2	6	5
3	6	5
4	6	5
5	7	6

(*b*)

Cell	*f*
3, 6	2
5, 4	2
1, 3	1
1, 5	1
2, 4	1
2, 6	1
4, 1	1
5, 7	1
6, 1	1
3, 2	1
4, 3	1
4, 6	1
5, 2	1
5, 3	1
7, 6	1

This is the end of the tree-search, $i=5$, and all nearest adjoining orders have been determined.

Short cuts There are a number of obvious short cuts which will occur to the reader during hand computations. For example, the essential result of Stage 1.1 for $i=5$ can be seen without actually carrying it out: first, changing 1,2 and 7,1 will break not six but five triads (since both changes break 1,2,7) leaving three so that at least one of the two changes still available must be of a cell of frequency two; secondly, changing 1,2 will reduce the frequency of cells 2,4, 4,1 and 6,1 from two to one (because of the triads 1,2,4 and 1,2,6) and changing 7,1 will similarly reduce the frequency of 1,3 and 1,5, so that there will be no cells of frequency two left. Again, Stage 1.4 can be shortened if it is noticed that none of the cells of frequency two will break the triad 1,2,7. However, such devices are difficult to systematize, and in any case hand computations are not really recommended.

3. DISCUSSION

Validity The validity of the procedure is easily seen. Any subset of cells, changing which will render the pair comparison table consistent must have the property mentioned under Step 2, namely that just one or two members of the subset occur in each circular triad. Since the procedure in effect examines all subsets (by examining all feasible classes of them) from the smallest upwards, until all minimal ones with this property are found, which also actually render the data consistent, it will necessarily determine i and all nearest adjoining orders.

Comparison with Remage and Thompson's algorithm A fairly florid example was necessary fully to display the procedure, which is easier to carry out than to describe, so that the reader may doubt whether it is any more efficient than the 399 computations needed with Remage and Thompson's algorithm.

In fact it does offer certain advantages. First, the length of the calculation for a fixed number of objects, m, appears to depend mainly upon i (a precise formulation of this relationship would probably be both very difficult to obtain and of little practical use) and for small i the procedure is very quick. The reader may verify for himself that the solution to Kendall's example (Table 8), where $i=2$, is evident at the end of step 2, or even step 1, after only a few minutes work, whereas Remage and Thompson's algorithm needs 156 computations. At the other extreme, a special pathological case, such as Table 9, where $i=7$, may be very lengthy (here each of the cells containing a plus sign occurs in just 2 of the 14 triads, so that the procedure reduces to trying out all $\binom{21}{7}=116{,}280$ feasible combinations, which is much worse even than the $7!=5{,}040$ computations required for complete enumeration). However, second, the procedure almost immediately obtains a minimum value for i, so that in practical applications hopelessly inconsistent data may be rejected without further ado, whereas Remage and Thompson's algorithm does not. Third, any

TABLE 8

	1	2	3	4	5	6
1	·	+	+	−	+	+
2	−	·	−	+	+	−
3	−	+	·	+	+	+
4	+	−	−	·	−	−
5	−	−	−	+	·	+
6	−	+	−	+	−	·

TABLE 9

	1	2	3	4	5	6	7
1	.	+	+	−	+	−	−
2	−	.	+	+	−	+	−
3	−	−	.	+	+	−	+
4	+	−	−	.	+	+	−
5	−	+	−	−	.	+	+
6	+	−	+	−	−	.	+
7	+	+	−	+	−	−	.

practical use of either method requires it to be embodied in a computer program and Remage and Thompson's algorithm appears awkward for this, since for the determination of the nearest adjoining orders the results of $m2^{m-1}-m^2$ (if matrix storage methods are used) or upwards of $2^{m+1}-1-(1-m^3)/(1-m)$ (if list processing is employed) computations must be simultaneously stored, which is rather extravagant of space. The present procedure can be conveniently implemented as a recursive procedure: such an implementation in the Elliott 803 dialect of ALGOL 60, has been developed, and copies may be obtained from the writer.

ACKNOWLEDGEMENTS

I am grateful to Professor R. Remage, Jr., Dr. P. Slater and Professor W. A. Thompson, Jr., for pointing out that some indication of the validity of the procedure was necessary, and to Mr. A. L. J. Wells and Mr. R. L. Hutchings for assistance in developing the ALGOL procedure.

REFERENCES

KENDALL, M. G. (1948). *Rank Correlation Methods*. London: Griffin.

REMAGE, R., Jr. and THOMPSON, W. A. (1966). Maximum likelihood paired comparison rankings. *Biometrika* **53**, 143–149.

PHILLIPS, J. P. N. (1967). Slater's *i* (and nearest adjoining orders) from pair comparisons or balanced incomplete block design rankings (computer abstract). *Br. J. math. statist. Psychol.* **20**, 261–262.

SLATER, P. (1960). The analysis of personal preferences. *Br. J. statist. Psychol.* **13**, 119–135.

SLATER, P. (1961). Inconsistencies in a schedule of paired comparisons. *Biometrika* **48**, 303–312.

Vol. 20
Part 2
227–238

The British Journal of Mathematical and Statistical Psychology

November
1967

THE EFFECT OF CRITERION SCORE GROUPING UPON ITEM PARAMETER ESTIMATION

By Frank B. Baker
University of Wisconsin

The differential effect of three grouping procedures upon the first two moments of the sampling distributions of the parameters X_{50} and β of the item characteristic curve were investigated empirically. The results of the study indicated that differences in the variances of $\hat{\beta}$ and $\hat{X}_{50}$ as well as in the bias of $\hat{X}_{50}$ could be attributed to differences in grouping procedures. Item parameter estimates possessing the minimum bias and variance within the study were those yielded by the grouping procedure which used equally spaced criterion score intervals and the same number of subjects within each interval.

1. Introduction

The assumption of a normal ogive relation between the probability of a subject answering an item correctly and his ability level underlies a number of classical item analysis procedures. Graphical techniques for estimating the parameters of location and scale of this ogive (the item characteristic curve) have been presented by Terman and Merrill (1937) and Das (1964); and analytical estimation techniques were developed by Ferguson (1942), Lawley (1943) and Finney (1944). That the logical basis of these procedures is the same as those used to estimate the median lethal dose in quantal-response bio-assay was shown by Finney (1944). In addition, the constant process used by Ferguson (1942) was independently derived in the bio-assay context by Berkson (1955) under the label of minimum normit chi-square.

Application of these estimation procedures to psychological test data usually involves grouping the criterion scores in some fashion. Typically, the criterion ability scale is divided into intervals of equal standard score width and the number of subjects falling in each interval is a function of the distribution of the underlying ability and of sample size. Each subject within an interval possesses a unique ability level, but due to grouping is considered to have attained the mid-point score. Tocher (1949) has shown that the effect of this type of grouping upon the estimates of the parameters of the item characteristic curve can be compensated by applying the usual Sheppards' corrections for grouping. A second grouping procedure, analogous to that followed in bio-assay, is one in which the criterion ability levels of interest are specified *a priori* and an equal number of subjects at each ability level are used. This procedure is widely used in analytical and empirical studies of the statistical properties of estimates of normal ogive parameters (see Berkson, 1955; Cramer, 1964) and can be used in routine item analysis when sufficient numbers of observations are available at each ability level. The latter would be typical of large-scale testing programmes where it is not economical to use all the data to estimate the item parameters. A third procedure (see Das, 1964) would be to partition the ability distribution into regions of equal frequency and use the mid-point of the region as the attained ability score of each subject within the interval.

A limited comparison of the first and third of these grouping procedures was performed by Das (1964) who compared estimates of the item parameters based upon a single data set having equal score intervals but an unequal number of subjects at each level with those estimates yielded by a second data set having unequal score intervals but an equal number of subjects at each level. On the basis of these *two* data sets she concluded that the estimates of the parameters of the item characteristic curve were " to a considerable extent independent of the basis for selecting class intervals, and hence, frequency, and range of ability within groups ".

Although the analytic work due to Tocher (1949) showed the effect of the typical item analysis grouping procedure upon item parameter estimates, the literature did not reveal a systematic investigation of the differential effect of various grouping procedures upon these estimates. Therefore, the present paper extends the previous work of Das (1964) and attempts to determine, through Monte Carlo procedures, if a differential effect exists.

2. PROCEDURES

The three grouping techniques employed were as follows:

(*a*) Equally spaced scores along the criterion ability scale were selected and an equal number of subjects having the specified score were used at each level—a procedure analogous to bio-assay practice.

(*b*) The ability scale was divided into equal standard score intervals and the frequency distribution of subjects was approximately normal. The mid-point score was used for all subjects in an interval—the typical item analysis procedure.

(*c*) The frequency distribution of subjects was normally distributed and divided into intervals of equal frequency. The ability score corresponding roughly to the centroid of an interval was used as the ability score of all subjects within the interval. The resulting centroid ability scores, though symmetric, were unevenly spaced about the mean—the Das procedure.

The numbers of grouping intervals investigated were seven and fifteen which were considered representative of usual item analysis practice. Both equal and unequal width grouping intervals were employed. The former involved interval widths of $0{\cdot}75\sigma$ and $0{\cdot}375\sigma$, for 7 and 15 intervals respectively. The latter employed interval widths which divided the normal density into 7 or 15 equal area intervals. Within each grouping procedure and number of intervals, three sample sizes were employed. The sample sizes were chosen to cover the range from approximately a single classroom to the point where large sample theory approximations hold for the variances of estimates of the item parameters (see Baker, 1962). These were 42, 84 and 168 for the seven-interval case; and 45, 90 and 180 for the fifteen-interval case. It was not possible to keep the sample size both small and equal across the two-interval specifications, hence the nearest possible sample sizes were used. When equal numbers of observations per grouping interval were employed, the number of observations was a multiple of 6 for the seven-interval case and of 3 for the fifteen-interval data. The observations in the unequal frequency data were roughly normally distributed over the $5{\cdot}25\sigma$ range of the grouping intervals.

The parameters of the item characteristic curve used were those adapted from bio-assay by Baker (1961) where:

X_{50} = the mean of the normal ogive, a measure of where on the ability scale the item functions;

β = the reciprocal of the standard deviation of the ogive, a measure of the item's power of discrimination.

In order to enhance the generality of the empirical results, nine sets of item parameter values, representing roughly high, medium, and low values of each parameter were employed.

The basic technique of the present paper was a Monte Carlo procedure in which the parameters of the item characteristic curve were used to determine the true proportions of correct response P_j at the specified ability levels. Given the true proportion of correct response P_j and the number of subjects, N_j, the observed number of correct responses, r_j, at that ability level was generated. The generated r_j were random and followed a binomial distribution having parameters P_j and N_j. The observed proportion of correct response $p_j = r_j/N_j$ was generated for each ability level and provided the data upon which estimates of the item parameters X_{50} and β were based. The minimum normit chi-square estimation procedure (Berkson, 1955) was used because it was non-iterative and therefore minimized the amount of expensive computer time. The generation-estimation procedure was replicated 100 times for each data set and item and parameter combination.

3. RESULTS

Seven Intervals

The empirical results based upon the seven-interval data are presented in Table 1 and the results it contains pertaining to X_{50} and β are discussed separately. For the specification $\beta = 0{\cdot}10$ and over all levels of X_{50} the three grouping procedures yielded sampling distributions of $\hat{\beta}$ having approximately the same mean. The magnitude of the standard deviations of the sampling distributions of $\hat{\beta}$ readily orders the three grouping procedures. For a sample size of 168 the order was: equal interval and equal frequency (0·06), equal interval and unequal frequency (0·09) and unequal interval and equal frequency (0·12).

TABLE 1. SUMMARY STATISTICS OF SAMPLING DISTRIBUTIONS OF ITEM PARAMETER ESTIMATES BASED UPON SEVEN ABILITY LEVELS

Item Parameter		Summary Statistics	Equal Interval						Unequal Interval		
			Equal Frequency			Unequal Frequency			Equal Frequency		
X_{50}	β		$N=42$	$N=84$	$N=168$	$N=42$	$N=84$	$N=168$	$N=42$	$N=84$	$N=168$
−1·5	0·10	$\hat{\beta}$	0·11	0·10	0·09	0·07	0·09	0·11	0·09	0·08	0·09
		$S_{\hat{\beta}}$	0·11	0·09	0·07	0·14	0·13	0·09	0·20	0·17	0·12
		$\hat{X}_{50}$	−2·85	−1·40	−2·81	−2·23	0·90	−1·32	*	−0·59	0·25
		$S_{\hat{X}_{50}}$	15·52	7·24	5·56	19·40	25·30	4·52	*	5·18	14·11
−1·5	0·40	$\hat{\beta}$	0·35	0·37	0·39	0·25	0·32	0·36	0·36	0·38	0·40
		$S_{\hat{\beta}}$	0·11	0·09	0·07	0·11	0·12	0·10	0·23	0·19	0·13
		$\hat{X}_{50}$	−1·60	−1·58	−1·54	−3·00	−1·86	−1·77	−2·53	−2·20	−1·77
		$S_{\hat{X}_{50}}$	1·14	0·56	0·42	3·20	2·21	0·75	22·02	4·57	1·30
−1·5	0·70	$\hat{\beta}$	0·44	0·55	0·63	0·34	0·45	0·55	0·47	0·58	0·64
		$S_{\hat{\beta}}$	0·10	0·08	0·07	0·10	0·11	0·10	0·19	0·19	0·14
		$\hat{X}_{50}$	−1·62	−1·57	−1·50	−2·59	−2·07	−1·78	−2·30	−2·29	−1·64
		$S_{\hat{X}_{50}}$	0·69	0·37	0·23	1·57	0·89	0·39	1·62	4·31	0·43
0·0	0·10	$\hat{\beta}$	0·11	0·10	0·10	0·07	0·12	0·09	0·12	0·12	0·09
		$S_{\hat{\beta}}$	0·13	0·10	0·06	0·13	0·13	0·09	0·23	0·17	0·12
		$\hat{X}_{50}$	8·45	0·36	−0·77	*	−1·93	−2·03	−0·53	4·89	−1·56
		$S_{\hat{X}_{50}}$	73·41	5·99	5·02	*	22·87	13·56	14·21	44·95	10·50
0·0	0·40	$\hat{\beta}$	0·37	0·40	0·41	0·30	0·34	0·37	0·41	0·38	0·38
		$S_{\hat{\beta}}$	0·11	0·11	0·08	0·13	0·12	0·09	0·22	0·19	0·13
		$\hat{X}_{50}$	−0·02	0·09	−0·05	−0·04	−0·06	0·04	0·21	0·07	−0·05
		$S_{\hat{X}_{50}}$	0·56	0·44	0·26	0·99	0·53	0·25	1·30	1·20	0·33

Table 1 (*Continued*)

Item Parameter		Summary Statistics	Equal Interval						Unequal Interval		
			Equal Frequency			Unequal Frequency			Equal Frequency		
X_{50}	β		$N=42$	$N=84$	$N=168$	$N=42$	$N=84$	$N=168$	$N=42$	$N=84$	$N=168$
0·0	0·70	$\hat{\beta}$	0·57	0·66	0·69	0·43	0·57	0·64	0·63	0·72	0·70
		$S_{\hat{\beta}}$	0·09	0·11	0·08	0·11	0·11	0·09	0·22	0·18	0·12
		$\hat{X}_{50}$	0·02	−0·00	−0·01	−0·02	−0·04	0·02	0·02	0·01	0·01
		$S_{\hat{X}_{50}}$	0·29	0·24	0·19	0·39	0·26	0·16	0·37	0·21	0·15
+1·5	0·10	$\hat{\beta}$	0·11	0·09	0·10	0·08	0·11	0·10	0·10	0·11	0·10
		$S_{\hat{\beta}}$	0·13	0·09	0·06	0·13	0·12	0·10	0·26	0·17	0·11
		$\hat{X}_{50}$	0·62	−7·52	0·43	*	35·34	7·48	3·29	−0·69	1·99
		$S_{\hat{X}_{50}}$	4·72	91·71	9·99	*	582·12	54·35	30·28	5·47	10·98
+1·5	0·40	$\hat{\beta}$	0·33	0·37	0·39	0·25	0·33	0·36	0·34	0·36	0·41
		$S_{\hat{\beta}}$	0·11	0·09	0·07	0·12	0·13	0·11	0·21	0·19	0·14
		$\hat{X}_{50}$	1·82	1·60	1·54	2·45	2·17	1·78	2·01	2·73	1·74
		$S_{\hat{X}_{50}}$	1·15	0·48	0·38	5·46	1·58	0·68	6·12	5·79	0·81
+1·5	0·70	$\hat{\beta}$	0·44	0·55	0·61	0·31	0·47	0·56	0·46	0·61	0·65
		$S_{\hat{\beta}}$	0·09	0·09	0·07	0·11	0·10	0·10	0·18	0·18	0·15
		$\hat{X}_{50}$	1·65	1·59	1·52	3·00	1·98	1·79	2·89	2·04	1·63
		$S_{\hat{X}_{50}}$	0·60	0·40	0·24	1·79	0·68	0·39	3·93	2·86	0·39

* Excessively large values not reported.

It should be noted that the standard deviation of the last procedure is twice that of the first. The sampling distributions of $\hat{\beta}$ for $\beta = 0{\cdot}40$ had a minimum bias under the equal interval and equal frequency procedure ($-0{\cdot}07$ to $0{\cdot}01$) and a maximum bias under the equal interval and unequal frequency procedure ($-0{\cdot}15$ to $-0{\cdot}01$). Increasing sample size from $N = 42$ to $N = 168$ reduced the bias to very small quantities ($-0{\cdot}02$ to $+0{\cdot}02$) under all three grouping procedures. The minimum standard deviation of $\hat{\beta}$, for $N = 168$, was achieved under the equal interval and equal frequency procedure ($0{\cdot}07$) and the maximum ($0{\cdot}13$) under the unequal interval and equal frequency procedure.

Increasing the level of discrimination to $\beta = 0{\cdot}70$ resulted in a consistent underestimate of β by the mean $\hat{\beta}$. At the smallest sample size the mean $\hat{\beta}$ ranged from $0{\cdot}31$ to $0{\cdot}63$ over all levels of X_{50} and grouping procedures. Increasing sample size to $N = 168$ resulted in values closer to nominal but none the less smaller than $0{\cdot}70$ (approximately $0{\cdot}63$). The standard deviation of $\hat{\beta}$ was greatest over all sample sizes and values of X_{50} for the unequal interval and equal frequency procedure ($0{\cdot}12$ to $0{\cdot}22$) and least for the equal interval and equal frequency procedure ($0{\cdot}07$ to $0{\cdot}11$). Again, the variability of $\hat{\beta}$ under the former was approximately twice that under the latter procedure.

In general, the mean of the sampling distribution of $\hat{\beta}$ was a reasonable estimate of β when values of $\beta = 0{\cdot}10$ and $0{\cdot}40$ were involved, but all three grouping procedures underestimated the population value of $\beta = 0{\cdot}70$. The sharpest contrast among grouping methods was evident in the standard deviations of $\hat{\beta}$. The unequal interval and equal frequency procedure consistently yielded sampling distributions of $\hat{\beta}$ which were twice as variable as those from the equal interval and equal frequency procedure.

In that X_{50} is a function of the reciprocal of β ($X_{50} = -\alpha/\beta$), the sampling properties of $\hat{X}_{50}$ were influenced by the value of β. When β had a value of $0{\cdot}10$ the random data generator produced many samples which yielded values of $\hat{\beta}$ near zero. The occurrence of such values resulted in estimates of X_{50} which were extremely large and the corresponding variance of the sampling distribution of $\hat{X}_{50}$ became unmanageably large. Because of this, several cells in Table 1 have no entries and others contain extremely large values of $S_{\hat{X}_{50}}$. At the higher levels of β the generation of such data sets rarely occurs and the sampling distributions of $\hat{X}_{50}$ were reasonable approximations to their known large sample distributions.

In that the data generated from the specification of $X_{50} = -1{\cdot}50$ and $+1{\cdot}50$ are used to estimate the parameters of the same normal ogive from its upper and lower portions respectively, the observed similarity of the sampling statistics at these two levels suggests that adequate Monte Carlo procedures were employed. When $\beta = 0{\cdot}40$ or $0{\cdot}70$, the population values of $X_{50} = -1{\cdot}50$ and $+1{\cdot}50$ were closely estimated by the mean value of $\hat{X}_{50}$ yielded by the equal interval and equal frequency and the unequal interval and equal frequency grouping procedures. The variability of the estimates produced by the latter procedure was usually about twice as large as observed for the former. The equal interval and unequal frequency procedure yielded mean values of $\hat{X}_{50}$ which had a bias of at least one-half a standard deviation unit. This bias is sufficiently large to affect one's interpretation of an item. The variability of the $\hat{X}_{50}$ yielded by this procedure was similar to that observed in the unequal interval and equal frequency results.

At the middle level, $X_{50}=0{\cdot}0$, with β at 0·40 or 0·70, the mean of the sampling distribution of $\hat{X}_{50}$ was a very close estimate (0·01 to 0·09) of X_{50} for all three grouping procedures. In addition, the standard deviations were roughly the same for all three grouping procedures.

The estimates of X_{50} yielded over all data sets by the equal interval and equal frequency grouping procedure possessed the least bias and the minimum variability of the three grouping procedures. It should be noted, however, that when $\beta=0{\cdot}10$, the sampling properties of $\hat{X}_{50}$ were consistently undesirable over all grouping procedures and data sets.

Fifteen Intervals

The results based upon the data for 15 intervals are presented in Table 2 and the relationships among the grouping procedures are similar to those observed in Table 1. Several characteristics of the data appear to be related to the increased number of intervals. The mean value of $\hat{\beta}$ ($N=180$) tended to underestimate the value of $\beta=0{\cdot}70$ to a slightly greater degree than was evidenced in the seven-interval data. The estimates had a mean of approximately 0·55 in the fifteen-interval results and of about 0·63 in the seven-interval results. The most discernible effect of increased number of grouping intervals upon the sampling properties of $\hat{X}_{50}$ was in the reduction in the occurrence of unrealistically large values of $\hat{X}_{50}$ when $\beta=0{\cdot}10$. In Table 1 there were three cells in which no $\hat{X}_{50}$ values were reported and in Table 2 there were none. The final effect attributable to increased number of intervals can be seen in the relationship of sample size to the summary statistics of $\hat{\beta}$. When 7 intervals were used, both the bias and variability of $\hat{\beta}$ decreased rapidly as sample size increased. However, when 15 intervals were employed the bias and variability of $\hat{\beta}$ changed very little as sample size increased. The initial values of both the bias and variability were less in the fifteen-interval results than in the seven-interval data.

4. DISCUSSION

Inspection of both the grouping procedures and the normal equations solved by the minimum normit chi-square estimation technique reveals a number of factors of importance in the present study. Fundamentally, the estimation technique is a weighted least-squares procedure which fits a linear regression line to the normal deviate transformations of the observed proportions of correct

TABLE 2. SUMMARY STATISTICS OF SAMPLING DISTRIBUTION OF THE PARAMETER ESTIMATES BASED UPON FIFTEEN ABILITY LEVELS

Item Parameter		Summary Statistics	Equal Interval						Unequal Interval		
			Equal Frequency			Unequal Frequency			Equal Frequency		
X_{50}	β		$N=45$	$N=90$	$N=180$	$N=45$	$N=90$	$N=180$	$N=45$	$N=90$	$N=180$
−1·5	0·10	$\hat{\beta}$	0·08	0·09	0·09	0·03	0·06	0·08	0·07	0·09	0·11
		$S_{\hat{\beta}}$	0·09	0·08	0·06	0·08	0·09	0·08	0·16	0·12	0·09
		$\hat{X}_{50}$	−2·58	0·74	−2·72	−2·71	−1·80	0·42	−0·88	−0·37	+1·48
		$S_{\hat{X}_{50}}$	28·18	16·71	10·72	17·36	22·16	11·24	10·14	9·84	30·17
−1·5	0·40	$\hat{\beta}$	0·24	0·33	0·37	0·11	0·20	0·30	0·23	0·35	0·37
		$S_{\hat{\beta}}$	0·07	0·08	0·06	0·07	0·08	0·08	0·13	0·14	0·10
		$\hat{X}_{50}$	−1·86	−1·65	−1·55	−4·34	−2·33	−1·98	−0·55	−2·54	−1·69
		$S_{\hat{X}_{50}}$	1·65	0·84	0·32	26·81	7·60	0·84	10·76	6·35	0·72
−1·5	0·70	$\hat{\beta}$	0·30	0·44	0·52	0·16	0·27	0·41	0·31	0·49	0·56
		$S_{\hat{\beta}}$	0·07	0·05	0·05	0·05	0·08	0·07	0·12	0·12	0·10
		$\hat{X}_{50}$	−1·75	−1·53	−1·52	−3·76	−3·31	−2·14	−2·48	−1·76	−1·67
		$S_{\hat{X}_{50}}$	0·80	0·36	0·25	2·23	3·15	0·46	2·28	0·77	0·34
0·0	0·10	$\hat{\beta}$	0·07	0·10	0·11	0·03	0·07	0·08	0·06	0·10	0·10
		$S_{\hat{\beta}}$	0·08	0·07	0·06	0·08	0·10	0·09	0·15	0·12	0·10
		$\hat{X}_{50}$	−4·28	−0·15	0·15	−0·60	1·58	−0·78	−0·39	−0·22	−0·37
		$S_{\hat{X}_{50}}$	38·20	4·28	1·71	13·77	25·80	8·04	3·09	3·56	2·42
0·0	0·40	$\hat{\beta}$	0·28	0·35	0·39	0·15	0·22	0·32	0·31	0·38	0·39
		$S_{\hat{\beta}}$	0·08	0·08	0·06	0·07	0·09	0·08	0·14	0·14	0·12
		$\hat{X}_{50}$	−0·02	0·00	−0·03	−0·23	−0·27	−0·01	−0·34	0·05	0·01
		$S_{\hat{X}_{50}}$	0·51	0·37	0·26	2·69	2·04	0·34	2·38	0·48	0·27

TABLE 2 (*Continued*)

Item Parameter		Summary Statistics	Equal Interval						Unequal Interval		
			Equal Frequency			Unequal Frequency			Equal Frequency		
X_{50}	β		$N=45$	$N=90$	$N=180$	$N=45$	$N=90$	$N=180$	$N=45$	$N=90$	$N=180$
0·0	0·70	$\hat{\beta}$	0·40	0·54	0·63	0·23	0·37	0·52	0·46	0·62	0·64
		$S_{\hat{\beta}}$	0·05	0·06	0·04	0·07	0·08	0·07	0·11	0·13	0·11
		$\hat{X}_{50}$	0·01	0·03	−0·00	−0·66	0·02	−0·02	0·03	−0·03	0·02
		$S_{\hat{X}_{50}}$	0·33	0·21	0·16	6·84	0·33	0·16	0·32	0·24	0·14
+1·5	0·10	$\hat{\beta}$	0·07	0·08	0·10	0·05	0·07	0·07	0·08	0·09	0·08
		$S_{\hat{\beta}}$	0·09	0·08	0·06	0·09	0·09	0·09	0·15	0·13	0·11
		$\hat{X}_{50}$	−0·36	1·02	2·26	0·83	1·39	3·12	−1·12	0·43	1·61
		$S_{\hat{X}_{50}}$	13·61	5·49	3·82	4·05	32·02	27·52	11·26	14·10	8·42
+1·5	0·40	$\hat{\beta}$	0·25	0·33	0·37	0·12	0·20	0·29	0·26	0·34	0·38
		$S_{\hat{\beta}}$	0·08	0·07	0·06	0·06	0·08	0·08	0·14	0·15	0·10
		$\hat{X}_{50}$	2·05	1·58	1·50	3·16	3·53	2·03	2·53	1·84	1·61
		$S_{\hat{X}_{50}}$	3·85	0·57	0·29	11·24	4·04	0·69	3·66	1·70	0·63
+1·5	0·70	$\hat{\beta}$	0·30	0·45	0·54	0·26	0·16	0·42	0·33	0·46	0·58
		$S_{\hat{\beta}}$	0·06	0·06	0·05	0·07	0·06	0·07	0·13	0·11	0·10
		$\hat{X}_{50}$	1·84	1·52	1·55	3·45	4·77	2·13	2·08	1·90	1·66
		$S_{\hat{X}_{50}}$	1·08	0·42	0·24	2·54	6·04	0·46	3·05	0·76	0·33

response. Under such a regression model, the ability scores are assumed to be 'fixed'. Therefore the equal interval and equal frequency and unequal interval and equal frequency grouping procedures differ only in the location and spacing of the 'fixed' ability scores. Under the former procedure the ability scores were equally spaced over a $4{\cdot}5\sigma$ range and under the latter the ability scores were unequally spaced over a $2{\cdot}78\sigma$ range. The net effect was that in the unequal interval and equal frequency condition the data in the tail of the ability distribution is pooled with that much closer to the mean, effectively basing the item parameter estimates on a narrower range of ability. It also implies that only a segment of the item characteristic curve is fitted to a cluster of observed proportions possessing a rather limited range of values which reduces the effectiveness of the least squares paradigm. Under the equal interval and equal frequency procedure, a larger portion of the item characteristic curve is fitted to a wider range of observed proportions. This difference probably accounts for the considerably larger variability of both β and $\hat{X}_{50}$ under the unequal interval and equal frequency grouping procedure.

The summary statistics of the sampling distributions of the item parameters obtained under the equal interval and unequal frequency grouping procedure were intermediate to those of the other two procedures, with the exception of the large bias of $\hat{X}_{50}$ at the levels of $X_{50} = \pm 1{\cdot}5\sigma$. In that equal frequencies were used in the other two procedures, the unique properties of this procedure must be ascribed to the unequal number of observations at each ability score level. In both of the equal frequency procedures, the possible values of the observed proportions of correct response were the same at all ability levels, hence the differences between successive proportions were the same, i.e., the 'step-size' was constant. Such was not the case when each ability level involved a different number of observations, in that the possible values of p_j as well as the 'step-size' depend upon N_j. When small samples were involved, 1·0 and 0·0 were the only possible observed values at a number of levels in the tails of the ability distribution. The net effect is that the observed proportions of correct response near the mean of the ability distribution will be more smoothly distributed about the true P_j with a smaller 'step-size' than those in the tails. Although the observed proportions in the tails of the ability distribution are crude estimates of the true proportions, they also make a smaller contribution to the estimates of the item parameters. The weighting term in the normal equations is multiplied by N_j, thus a small N_j tends to reduce the contribution of a crude estimate of P_j. It should also be noted that under the minimum normit chi-square technique, observed values of $p_j = 1{\cdot}0$ or $0{\cdot}0$ are not admissible, so that one invokes the $1/2N$ rule (Berkson, 1955) replacing 0 by $1/2N_j$ and 1·0 by $1 - 1/2N_j$. Such a substitution tends also to reduce the value of $\hat{\beta}$ and its variability (Cramer, 1964). Because the weighting involves N_j, the equal interval and unequal frequency procedure tends to yield estimates which depend primarily upon those observed proportions near the mean of the ability distributions. Thus, the opportunity for one or more intervals to strongly influence the estimate exists and the possibility of a biased estimate is enhanced.

The role of ' step-size ' also makes itself apparent in the effect upon the means and variances of the item parameter estimates due to increasing sample size and to increasing the number of grouping intervals. When 7 intervals were employed, the bias of the mean of $\hat{\beta}$ and the variability of $\hat{\beta}$ decreased markedly when the number of observations per ability level increased from 6 to 24. Such a pronounced decrease was not observed in the fifteen-interval results where the summary statistics for $\hat{\beta}$ changed little when the number of observations per ability level increased from 3 to 12. When unequal frequencies were used with data based on 15 intervals, the ' step-size ' problem was aggravated, in that a large proportion (6/15, 8/15, and 8/15) of the ability levels had numbers of observations less than that employed under the equal frequency procedures. Consequently, the mean of $\hat{\beta}$ tended to underestimate β to a greater degree than that yielded by the equal interval and equal frequency procedure. This pattern of underestimation was also observed in the seven-interval data, but was more pronounced in the fifteen-interval data.

The empirical results cited above and their elaboration indicate that the ' best ' grouping procedure over all combinations of sample size, number of grouping intervals and item parameter values was the equal interval and equal frequency procedure. The next ' best ' was the equal interval and unequal frequency procedure which is analogous to typical item analysis practice. The sampling properties of the item parameter estimates yielded by the unequal interval and equal frequency grouping procedure suggested by Das (1964) were not as acceptable as those of either equal interval procedure. This procedure consistently yielded item parameter estimates having greater variability than either of the other grouping procedures.

The results of the present study suggest that discernible differences in the sampling properties of item parameter estimates can be attributed to the grouping procedure, i.e., the number and spacing of the ability scores used, as well as the number and distribution of the observations at each ability level. Operationally, the bias observed in the mean of $\hat{\beta}$ or $\hat{X}_{50}$ as a function of the grouping procedure was not great enough at the largest sample sizes to be of concern, with the exception of the bias of $\hat{X}_{50}$ under the equal interval and unequal frequency procedure. However, the variabilities of the estimates of $\hat{\beta}$ and $\hat{X}_{50}$ yielded by the unequal interval and equal frequency procedure were, in many cases, nearly twice as large as those yielded by the equal interval and equal frequency procedure.

The differences existing between grouping procedures were quite pronounced at the smallest sample size and, as would be expected, least pronounced at the largest sample size. Increasing sample size had the most effect within the seven-interval data, which suggests that the number of observations at each ability level is an important consideration. The overall similarity of the seven- and fifteen-interval data suggests that increasing the number of intervals and application of the $1/2N$ rule tends to compensate for the problems associated with large ' step-size '.

The FORTRAN 63 computer program used to generate and analyse the data of the present study is available from the author upon request.

REFERENCES

BAKER, F. B. (1961). Empirical comparison of the item parameters based on the logistic and normal functions. *Psychometrika* **26**, 235–246.

BAKER, F. B. (1962). *Empirical Determination of Sampling Distributions of Item Discrimination Indices and a Reliability Coefficient*. University of Wisconsin, Final Report U.S.O.E. Contract OE–2–10–071.

BERKSON, J. (1955). Estimate of the integrated normal curve by minimum normit chi square with particular reference to bio-assay. *J. Am. statist. Ass.* **50**, 529–549.

CRAMER, E. (1964). Some comparisons of methods of fitting the dosage response curve for small samples. *J. Am. statist. Ass.* **59**, 779–793.

DAS, R. S. (1964). Item analysis by probit and fractile graphical methods. *Br. J. statist. Psychol.* **27**, 51–64.

FERGUSON, G. A. (1942). Item selection by the constant process. *Psychometrika* **7**, 19–29.

FINNEY, D. J. (1944). The application of probit analysis to the results on mental tests. *Psychometrika* **9**, 31–39.

LAWLEY, D. N. (1943). On problems connected with item selection and test construction. *Proc. Roy. Soc. Edin.* **61A**, 273–287.

TERMAN, L. and MERRILL, MAUDE A. (1937). *Measuring Intelligence*. Boston: Houghton-Mifflin.

TOCHER, K. D. (1949). A note on the analysis of grouped probit data. *Biometrika* **36**, **9–17.**

NOTES AND CORRESPONDENCE

A NOTE ON A GENERATING ASSUMPTION IN McGEE'S MULTIDIMENSIONAL ANALYSIS OF 'ELASTIC' DISTANCES

By R. A. M. Gregson and P. N. Russell
University of Canterbury, New Zealand

Evidence is offered that multidimensional similarity judgements do not behave in a way demanded for an elastic distance model advanced by McGee. Problems concerning the properties of similarity data which can be included in scaling models are indicated.

In a multidimensional scaling model differing in some important particulars from those already published, McGee (1966, p. 182) advances a psychophysical rationale for introducing a tolerance when mapping perceived similarities into a multidimensional space. He asserts that 'Psychological judgements which indicate relatively great separation of stimuli should be allowed greater error than judgements indicating close proximity. This conception is well documented in the copious literature on psychophysics and refers to the " resolving power " of the organism.' He again emphasizes the central role of this generating assumption in his model (p. 189) where he states the criterion . . . is explicitly rooted in a . . . time-honoured tradition that large disparities are less readily resolved than small disparities '. Whilst his method shows relatively little difference from others with which it is compared, he concludes ' a case can be made in favour of requiring a multidimensional scaling procedure to yield information concerning judgement variability '.

The problem here is that the psychophysical precedents which McGee uses as axiomatic may not be strictly relevant to the multidimensional situation. The unidimensional psychophysical situation, where judgements of stimulus intensity or stimulus magnitude are elicited, leads, in some cases, to a relation of the form

$$\sigma_x(x) = kx + b, \tag{1}$$

where $\sigma_x(x)$ is the variability of the physical correlate of stimuli judged equal to x. This form of equation with ψ (a magnitude response) for x was shown by Eisler (1963, 1965) to be a Weber function and may be the basis of a general psychophysical differential equation, with k corresponding to the Weber fraction.

When, however, we turn to comparison judgements in a multivariate stimulus context and propose to use these as data for a multidimensional scaling, the purely empirical question arises as to whether or not such judgements do, in fact, fit eqn. (1) with x or with ψ as the variable.

From records of recent gustatory psychophysics experiments we have conducted (Gregson, 1965, 1966 a, b, c; Gregson and Russell, 1965; Russell and Gregson, 1966), we have re-examined relations between mean δ_{ij} (McGee's notation here; but with δ_{ij} based on the perceived degree of difference, Δ_{ij}, or on a degree of confidence, C_{ij}, that two stimuli compared are the same, and not on direct estimates of similarity) and var δ_{ij} over the $M \times M$ cells of the complete paired comparison matrices which are the raw data in each multidimensional scaling we have attempted. In all such matrices we have made comparisons for δ_{ij}, δ_{ji}, and δ_{ii}, as it is in general true that

$$\delta_{ij} \neq \delta_{ji} \quad \text{and} \quad \delta_{ii} \neq 0,$$

and these inequalities are regular, reliable and psychologically meaningful. If a relation of the form of eqn. (1) obtains between δ_{ij} and var δ_{ij}, then it is to be expected that the

TABLE 1. RELATIONS BETWEEN PERCEIVED SIMILARITIES OF COMPLEX STIMULUS PAIRS AND THEIR VARIANCES FOR 13 MULTIDIMENSIONAL SCALING EXPERIMENTS

Reference	*Type of experiment*	*Response Type*	*N*	*No. of stimulus dimensions*	*Matrix size*	*r between cell means and variances*	*Source*
a	$M \times M'$	C	10	3	8× 8	−0·424	Gregson (1965)
b	$M \times M'$	C	12	4	10× 10	−0·53	Gregson (1966 a)
c	$M \times M$	Δ	60*	3	8× 8	+0·286	Russell and Gregson (1966)
d	$M \times M$	Δ	96*	4	10× 10	−0·113	Russell (unpubl.)
e	$M \times M'$	C	10	3+1 not taste	10× 10	−0·078	Gregson (1966 b)
f	$M \times M'$	C	16	4	5× 8	−0·566	Gregson (1966 c)
g	$M \times M'$	C	16	4	5× 8	−0·099	Gregson (1966 c)
h	$M \times M'$	C	16	4	5× 8	−0·399	Gregson (1966 c)
i	$M \times M'$	C	16	4	5× 8	−0·058	Gregson (1966 c)
j	$M' \times M'$	Δ	17	4	$\frac{1}{2}$(10× 10)	−0·404	Russell (unpubl.)
k	$M' \times M'$	Δ	18	4	$\frac{1}{2}$(10× 10)	−0·255	Russell (unpubl.)
l	$M' \times M'$	Δ	16	4	$\frac{1}{2}$(10× 10)	−0·077	Russell (unpubl.)
m	$M' \times M'$	Δ	16	4	$\frac{1}{2}$(10× 10)	−0·087	Russell (unpubl.)

* Incompletely replicated designs.

correlation $r(\delta_{ij}, \text{var } \delta_{ij})$ over $M \times M$ cells with $(M \times M - 2)$ d.f. would be significantly positive because r measures a linear component of regression and eqn. (1) is linear.

Three sorts of experiments were involved, intramodal ($M \times M$ or $M' \times M'$), and intermodal ($M \times M'$). In the first case either taste mixtures, M, were compared with one another; or histograms, M', were so compared. In the intermodal case, taste mixtures were compared with histogram pictures of taste mixtures. Table 1 above sets out: the type of experiment; the response type, C_{ij} or Δ_{ij}; the number of subjects involved, N; the number of stimulus dimensions introduced; the matrix size; and the correlation between δ_{ij} and var δ_{ij}, where δ_{ij} is taken as the mean Likert scale value of a difference or matching task pooled over all subjects. This is important, as replicated scalings on a single subject may behave differently; though in the case of gustatory psychophysics we strongly doubt this, and this empirically supported doubt has been a reason for the use of data pooled over subjects. The original study is also given where relevant data have been published, although these correlations may not have been reported.

Further examination of the plots of $\delta_{ij}/\text{var } \delta_{ij}$ indicated that some of these correlations are numerically small, partly because the relationship in question is curvilinear; in these cases the largest variances are associated with intermediate similarity magnitudes and the smallest variances with large or small δ_{ij}; that is, with stimulus mixture pairs which are generally perceived as very similar or very dissimilar. So, on the available evidence, there exist at least two types of experiment where McGee's axiom, and hence his model, are invalid mappings of behaviour.

Some more light may be cast on this problem by consideration of cases (*f*) and (*h*) as against (*g*) and (*i*) in Table 1. The larger negative values are associated with a task in which subjects were paced to perform relatively quickly, whereas the smaller values are associated with precisely the same stimulus task performed more slowly. However, in cases (*j*) and (*k*) the task was one in which 10 sec were allowed for a response, whilst for (*l*) and (*m*) the same task had to be performed in 2 sec. Also, in cases (*k*) and (*m*), anchor stimulus comparisons illustrating large and small distances, as predicted from Kruskal's (1964) formula, were provided for subjects at the start of the task. It seems pertinent to observe that subjective similarity is a metric that does not necessarily behave like distance or magnitude, for similarity is closed, bounded at both ends, where either stimulus pairs are identical or they are as dissimilar as they can be. This is not an analogue of stimulus magnitude which is bounded by a lower and an upper threshold unless stimuli used in a particular experiment cover a much wider range than is usually practicable.

Further complications arise because: (i) identity is single-valued but maximum dissimilarity is multi-valued; (ii) with complex stimuli their interdimensional variance or internal redundancy affects the reliability with which they are identified, which in turn affects the variance of the perception of δ_{ij} in which they are involved; (iii) there is inadequate evidence on the relation between the response variables similarity and dissimilarity—if $\sum_{ij}$ is a similarity response, Δ_{ij}, a dissimilarity response and δ_{ij} the psychological distance which they are intended to estimate, presumably $\Delta_{ij} = k'\delta_{ij}$, but it is not clear if $\sum_{ij} \Leftrightarrow k'(1 - \delta_{ij})$ or $\sum_{ij} \Leftrightarrow k'(1/\delta_{ij})$, or if $\sum_{ij} = (\Delta_{\max} - \Delta_{ij})$, or with what generality these relations might hold. Unpublished data collected by Russell suggest that perceived dissimilarity is linearly related to the degree of subjective confidence of a difference, C_{ij}, in part because we have scored $C_{ij} \Leftrightarrow k'\delta_{ij}$, but that similarity is not necessarily linearly related to dissimilarity.

We agree with McGee when he concludes ' experience appears to be the teacher in the use of multidimensional scaling procedures ' and think that only by a closer examination of the properties of the judgements and scales required to generate data for multidimensional similarities can we properly decide what the appropriate axioms might be. *A priori* McGee's axiom was as good as many others that have been advanced by workers in this field; *a posteriori* his model seems to be restricted to experimental situations in which the data have properties which we have not encountered. As no such restriction is imposed by

S.P.

Kruskal's (1964) methods, then if we are not prepared to commit ourselves to the most plausible detailed structure of our distance measure on the grounds of evidence collected in closely pertinent psychophysical situations, it would seem best to use non-metric rather than elastic spaces. If McGee is prepared to try different axioms to eqn. (1) his model could still be elastic and have more general validity.

REFERENCES

EISLER, H. (1963). A general differential equation in psychophysics: Derivation and empirical test. *Rep. Psychol. Lab., Univ. Stockholm*, No. 160.

EISLER, H. (1965). On psychophysics in general and the general psychophysical differential equation in particular. *Rep. Psychol. Lab., Univ. Stockholm*, No. 195.

GREGSON, R. A. M. (1965). Representation of taste mixture cross-modal matching in a Minkowski r-metric. *Aust. J. Psychol.* **17**, 195–204.

GREGSON, R. A. M. (1966 a). Theoretical and empirical multidimensional scalings of taste mixture matchings. *Br. J. math. statist. Psychol.* **18**, 59–75.

GREGSON, R. A. M. (1966 b). *Scaling Perceived Similarities between Complex Tastes which include Ethanol.* (Univ. Canterbury, Dept. Psychol. and Sociol. Res. Project 9.)

GREGSON, R. A. M. (1966 c). Cross-modal matching of histograms and four-component taste mixtures, with one component fixed, under two pacing conditions. *Percept. mot. Skills* **23**, 183–190.

GREGSON, R. A. M. and RUSSELL, P. N. (1965). *Problems and Results in the Scaling of Intermodal and Intramodal Complex Taste Similarities by D* Metrics.* (Univ. Canterbury, Dept. Psychol. and Sociol. Res. Project 8.)

KRUSKAL, J. B. (1964). Nonmetric multidimensional scaling: A numerical method. *Psychometrika* **29**, 115–129.

MCGEE, V. E. (1966). The multidimensional analysis of 'elastic' distances. *Br. J. math. statist. Psychol.* **19**, 181–196.

RUSSELL, P. N. and GREGSON, R. A. M. (1966). A comparison of intermodal and intramodal methods in the multidimensional scaling of 3-component taste mixtures. *Aust. J. Psychol.* **18**, 244–254.

A REPLY TO SOME CRITICISMS OF ELASTIC MULTIDIMENSIONAL SCALING

By Victor E. McGee
Dartmouth College, Hanover, New Hampshire

Gregson and Russell (1967) have raised an important issue in connection with the error function for dissimilarity data. In McGee (1965, 1966 a, b) the discrepancy between d_{ij} and $\bar{d}_{ij}$ was assumed to be normally distributed with zero mean and standard deviation proportional to $\bar{d}_{ij}$ (or d_{ij}, depending on which of d_{ij} or $\bar{d}_{ij}$ one wishes to call the random variable). This type of error function is a direct outcome of requiring Weber's law to hold for dissimilarity judgements and so we shall call it the Weber function. Gregson and Russell take issue with the validity of this error function in the context of experiments with taste stimuli and their criticism is well taken. Since a number of other investigators have given me the benefit of their reactions to 'elastic' multidimensional scaling (called EMD scaling from now on) and since they point to the need for clarification of this same issue and one or two others, this reply will be organized as follows. First, the form of Gregson and Russell's criticism will be examined with a view to showing that, although the shape of the error function needs to be examined empirically (as they say), their reasons for rejecting the Weber function are not compelling. Second, it has been argued that making $\lambda_{ij}d_{ij}$ constant for all 'interpoint springs' is 'completely arbitrary'. It will be shown that if Weber's law is assumed to hold for dissimilarity judgements then $\lambda_{ij}\bar{d}_{ij}=c^2$ is a necessary consequence. Third, making use of some valuable advice from Carroll it will be shown that the EMD procedure can be made quite general. In particular, it can be made to handle any error function desired. And fourth, the matter of the chi-square decision rule will be clarified.

(*a*) *A Reply to Gregson and Russell*

These two investigators point to 'the purely empirical question . . . as to whether or not such [dissimilarity] judgements do, in fact, fit' some specified error function. This is a good question. They go on to report a number of experiments using taste mixtures and conclude that 'the largest variances are associated with intermediate similarity magnitudes and the smallest variances with large or small δ_{ij}'. Their statements are somewhat too emphatic in my opinion. In attempting to verify the Weber function as used in EMD scaling, it was found to be well-nigh impossible to analyse categorical data in a reasonable manner for the purpose of estimating var δ_{ij}. The trouble is that categorical data represent a reduction of information assumed to exist on some underlying continuum. It is not possible to recapture the information lost in the process of categorizing. Hence var δ_{ij}, where δ_{ij} is a Likert rating for a category, is not a satisfactory estimate of what we are looking for. As Kruskal says of Shepard's criterion, 'we shall find ourselves doing arithmetic with dissimilarities'. As one reviewer has put it, 'once the best monotone transformation enters into the picture, neither index (McGee's *W* or Kruskal's stress) tells one much about the variances of judgements on the continuum *on which the judgements were made*'. It is difficult to see how any useful statement can be made about variances on a categorical scale whose metric is unknown. (Parenthetically, it is easy to show that equal variance distributions on a continuum between zero and some finite upper limit show up as larger variances in intermediate categories and smaller variances at the lower and upper ends when the continuum is chopped into equal-interval categories.)

With respect to the inequalities $\delta_{ij}\neq\delta_{ji}$ and $\delta_{ii}\neq 0$, I have found in work with distorted speech samples (which must be presented successively) that good judges (i.e. those whose correlation between two complete sets of 105 dissimilarity judgements over the same pairs was significantly higher than 0·5) do not show the first inequality consistently. Worse judges do consistently show order effects, strongly in favour of the second speech sample. In the case of the second inequality, since we do not allow a zero-judgement, it is necessary to determine what *non*-zero scale value represents the zero point.

In sum, I agree that empirical work on judgement inequalities and judgement variability needs to be done. Such work fits the motivation for developing the EMD procedure; namely, that multidimensional scaling procedures themselves should 'yield information concerning judgement variability'. If and when an experimenter obtains information about his error function, then the general EMD procedure to be explained in (*c*) below may be applied.

(*b*) *Given the Weber Function,* $\lambda_{ij}d_{ij}=c^2$ *is true*

Let us assume that Weber's law holds in the case of dissimilarity data. Then the following proof (due to Joseph W. Serene) shows that $\lambda_{ij}d_{ij}=c^2$ is a necessary consequence.

The force required to distort the spring, d_{ij}, to a new length, d_{ij}, is given by

$$F_{ij}=k_{ij}(d_{ij}-d_{ij}).$$

The work needed to effect this distortion is found by integrating with respect to distance:

$$w_{ij}=\int_{d_{ij}}^{d_{ij}} F_{ij} = \int k_{ij}(d_{ij}-d_{ij}) = \frac{k_{ij}}{2}(d_{ij}-d_{ij})^2. \tag{1}$$

The work criterion then has the form:

$$W=\sum_i \sum_j \frac{k_{ij}}{2}\ (d_{ij}-d_{ij})^2.$$

In terms of eqn. (1) in McGee (1966 a) k_{ij} and λ_{ij} are related by:

$$\lambda_{ij}=k_{ij}d_{ij}/2.$$

The k_{ij}'s depend upon the stiffness of the springs, which in turn depends upon the tolerance for error in the data. Assuming Weber's law to hold, the k_{ij}'s should be chosen so that whenever

$$\frac{(d_{ij}-d_{ij})}{d_{ij}} = \frac{(d_{mn}-d_{mn})}{d_{mn}}, \tag{2}$$

then

$$w_{ij} = w_{mn}.$$

By the definition of the work criterion (eqn. (1) here),

$$\frac{k_{ij}}{2}(d_{ij}-d_{ij})^2 = \frac{k_{mn}}{2}\ (d_{mn}-d_{mn})^2,$$

$$\frac{k_{ij}}{k_{mn}} = \frac{(d_{mn}-d_{mn})^2}{(d_{ij}-d_{ij})^2}.$$

Using eqn. (2), this can be rewritten:

$$\frac{k_{ij}}{k_{mn}} = \frac{d_{mn}{}^2}{d_{ij}{}^2},$$

which is satisfied if

$$k_{ij}' = C/d_{ij}{}^2. \tag{3}$$

The work criterion then becomes

$$W=\sum_i \sum_j \frac{C}{2}\left(\frac{d_{ij}-\bar{d}_{ij}}{\bar{d}_{ij}}\right)^2 . \tag{4}$$

Since the magnitude of the constant C has no significance as yet, eqn. (4) reduces immediately to eqn. (4) in McGee (1966); namely,

$$W=\sum_i \sum_j (c\, p_{ij})^2,$$

where $p_{ij}=(d_{ij}-\bar{d}_{ij})/\bar{d}_{ij}$.

The above proof establishes the necessity of the relationship $\lambda_{ij}\bar{d}_{ij}=c^2$ if Weber's law is assumed to hold. Without wanting to appear fixated, it should be noted that several investigators feel a natural empathy for this assumption even in connection with similarity data: for example, Jones (1959, 1967).

(*c*) *Generalizing the EMD Procedure*

In a particularly fruitful series of letters starting in 1964, J. D. Carroll has contributed much to my thinking about EMD methodology. The following generalization is due largely to him.

Consider a set of dissimilarities which are monotonically transformed into distances, $\{\bar{d}_{ij}\}$. We shall call these the observed distances. Then the EMD model can be stated formally as follows:

$$\bar{d}_{ij}=d_{ij}+e_{ij}, \tag{5}$$

where d_{ij} is the ' true ' distance between i and j, and e_{ij} is an error term which is assumed normally distributed with mean zero and variance σ_{ij}^2. We further allow that this error term is functionally related to the true distances:

$$\sigma_{ij}^2=f(d_{ij}), \tag{6}$$

where f is any function, monotonic or non-monotonic. The work function to be minimized is:

$$W=\sum_i \sum_j c_{ij}(d_{ij}-\bar{d}_{ij})^2, \tag{7}$$

where

$$c_{ij}=C/\sigma_{ij}^2 \quad (C>0). \tag{8}$$

Combining eqns. (6), (7) and (8), we obtain:

$$W=C\sum_i \sum_j \frac{(d_{ij}-\bar{d}_{ij})^2}{f(d_{ij})} . \tag{9}$$

Given the prior assumptions (eqns. (5) and (6)), this work function is immediately distributed as $C\chi^2$ with certain degrees of freedom. (The degrees of freedom problem is treated in the next section.) The constant C can be interpreted as the ratio of error variance to true variance.

The general model shown above has the advantage of leaving the choice of the error function to the experimenter. However, as a note of caution and by way of correcting an error in the 1965 FORTRAN program, it should be noted that for each error function it is necessary to determine a specific algorithm (subroutines ALG and AMALG in the FORTRAN program) for finding the best monotone transforms of the dissimilarities. The use of the Kruskal's algorithm in McGee (1965, 1966 a) was inappropriate for the EMD procedure. In the latest version of the computer program (Spring 1967) subroutines ALG and AMALG are the appropriate ones for the Weber function in EMD. Copies of the improved program may be obtained from the author.

In response to Gregson and Russell's final statement: ' If McGee is prepared to try different [generating] axioms . . . his model could still be elastic and have more general validity ', I believe this generalization above satisfies their suggestion.

(*d*) *The Chi-square Criterion*

Criticism of the use of the chi-square criterion has ranged from ' the interpretation of the work criterion as a chi-square test is wrong . . . the chi-square value is meaningless as a statistic ' to ' the use of the chi-square distribution seems very clever and very sensible: only it should be labelled an approximation '. In addition to these comments, Carroll has insisted from the beginning that the degrees of freedom, as indicated in McGee (1965, 1966 a), are computed incorrectly. My original formulation stated that, for an r-space fit,

$$\text{d.f. for } r\text{-space fit} = \binom{k}{2} - \binom{k-r}{2}. \tag{10}$$

Carroll argues that the formula should be:

$$\text{d.f. for } r\text{-space fit} = \binom{k-r}{2}. \tag{11}$$

In our more recent work we have decided that the appropriate formula should be:

$$\text{d.f. for } r\text{-space fit} = \binom{k}{2}. \tag{12}$$

There is a problem concerning how the numerical analysis should be interpreted. On the one hand, when discussing the EMD procedure, I argue in the following order: (i) obtain a set of dissimilarities, $\{\delta_{ij}\}$; (ii) transform these monotonically into distances, $\{\hat{d}_{ij}\}$; and (iii) find the best-fitting set $\{d_{ij}\}$ in the r-space solution. From this point of view we might consider the set $\{d_{ij}\}$ and the set $\{d_{ij}\}$ to be two independent sets of numbers. The work value needed to make the d_{ij}'s fit the $\hat{d}_{ij}$'s may then be interpreted as a chi-square value with degrees of freedom given by eqn. (12).

On the other hand, the numerical method proceeds in a different order: (i) obtain the set of dissimilarities, $\{\delta_{ij}\}$; (ii) obtain a set of distances $\{d_{ij}\}$ in r-space; and (iii) apply the ALG–AMALG subroutines to find the best-fitting monotone transforms, $\{d_{ij}\}$. From this point of view I believe Carroll's eqn. (11) is the appropriate one to use.

Clearly, the numerical procedure ties the d_{ij}'s and the d_{ij}'s together, so that they are not independent, but when it comes to specifying the shape of this functional dependence the analysis becomes intractable. In the current state of confusion I reject eqn. (10) and hold in abeyance the choice between eqns. (11) and (12). As a matter of fact, most users of EMD do not use the chi-square criterion in any explicit manner.

Prognosis

There is no doubt further room for improvement in the general EMD procedure. For example, the use of normal distributions is (to quote Carroll) ' not quite so compelling, especially when it is realized that negative distances make no sense, and thus some kind of "ceiling" (or "floor", in this case) must be operating. Intuitively, I would expect some sort of skewed (to the right) distribution rather than the normal '. However, it has been applied with success in a number of studies and the freedom from any one specific error function should make it more useful. Recent developments include the extension of EMD methodology to the study of reproducibility of multidimensional spaces (something akin to reliability analysis) and to hierarchical clustering of multidimensional spaces.

ACKNOWLEDGEMENTS

I should like to take this opportunity to express my appreciation to Drs. Gregson and Russell, Carroll, Jones, Kruskal and Shepard for their useful critiques.

REFERENCES

GREGSON, R. A. M. and RUSSELL, P. N. (1967). A note on a generating assumption in McGee's multidimensional analysis of 'elastic' distances. *Br. J. math. statist. Psychol.* **20,** 239–242.

JONES, LYLE V. (1959). Some invariant findings under the method of successive intervals. *Am. J. Psychol.* **72**, 210–220.

JONES, LYLE V. (1967). The invariance of zero-point scaling over changes in stimulus context. *Psychol. Bull.* (in press).

MCGEE, VICTOR E. (1965). Note on an 'elastic' multidimensional scaling procedure. *Percept. mot. Skills* **21**, 81–82.

MCGEE, VICTOR E. (1966 a). The multidimensional analysis of 'elastic' distances. *Br. J. math. statist. Psychol.* **19**, 181–196.

MCGEE, VICTOR E. (1966 b). Elastic multidimensional scaling : a hybrid technique. In *Proc. Int. Symp. Math. Computational Methods in Social Sciences, Rome.* (July 4–8, 1966.)

Vol. 20
Part 2
249–260

The British Journal of Mathematical and
Statistical Psychology

November
1967

COMPUTER SECTION

APL : A COMPUTING LANGUAGE DESIGNED FOR THE USER

By S. Hunka
University of Alberta

Traditionally computer languages have been thought of only in the context of communications with a computer. Most languages in common use today are still more dependent upon machine design characteristics than the needs of the user. APL, designed by K. E. Iverson, is a new language equally useful for purposes of communicating with a computer as it is for use in the classroom. This powerful language can be learned easily and applied to a wide range of problems.

1. Introduction

Because of the widespread need for students and researchers to solve their research problems with the assistance of a computer, it has been usual in the last decade to include computer programming as part of the university curriculum. In North America the programming language has typically been FORTRAN when the objective was to provide the student with a tool for solving problems and not an understanding of the internal operation of the computer itself.

Existing languages, although having marked advantages over symbolic codings such as MAP and FAP, still have one major disadvantage when used in the educational setting or where the computer is used as an aid to the solution of problems provided within the classroom. One of the major drawbacks rests with the length of time needed to instruct a researcher in the language, and even when the language is learned, frequently the researcher spends more time 'debugging' than upon understanding the concept embodied in the problem. A further criticism of existing computing languages is that they do not lend themselves to use in the classroom, particularly on the blackboard, for the description of algorithms.

Iverson (1962, 1966) has described a mathematical notation which has been adapted for computer usuage, and which can easily be learned and used in the classroom quite independently of a computer. The language, known as APL, is also perfectly suited for the description of algorithms.

A major advantage of APL resides in the fact that much of the symbolism specified is that which is customarily used in logic and mathematics. Thus the need for the user to master new symbols is removed. As an illustration, consider the simple problem which requires the addition and multiplication of the elements of a vector of order *n*.

Arithmetic Notation	*APL Notation*	*FORTRAN Notation*
Σx_i	$+/x$	SUM = 0 DO 1 J = 1,N 1 SUM = SUM+X(J)
Πx_i	$\times/x$	SUM = X(1) DO 1 J = 2,N 1 SUM = SUM*X(J)

The similarity between the Iverson notation and usual mathematical notation is striking. In fact, one might argue that the arithmetic notation is inadequate relative to that of Iverson because the Σ and Π do not convey to the reader the basic difference between the two arithmetic operations, i.e. that one involves an addition while the other a multiplication. Since the addition and multiplication operators are included as part of the Iverson notation for these operations, a clearer specification is made. In contrast to FORTRAN, the Iverson notation is obviously more meaningful and simpler.

All of the examples which follow have been run on an IBM 360/67 computer located at the University of Alberta. The interpreter used with this machine was provided through the courtesy of the T. J. Watson Research Laboratories, Yorktown Heights, New York. It should be emphasized that APL is being used with an interpreter rather than with a compiler. With the interpreter there is immediate execution of the instructions and the results are then available for display to the user. In our particular system, IBM 2741 terminals are used with a time-sharing system. One marked advantage being realized by the use of this system is that errors are immediately made evident by the system, and the researcher interacts with the computer in a ' conversational ' mode.

2. A Simple Description of the Iverson Notation

The only rule that is to be noted in the use of this new notation is that, although there is no hierarchy of operators, there is a right-to-left precedence of operations. An operator acts on the variables to its right up to a right parenthesis or the end of the statement. When an operator requires parameters on the right and left it may be assumed that the right-hand portion is calculated first and is related to that variable on the left by the operator. For purposes of clarity we shall define three types of operands and note these by the type used: scalars, x; vectors, $\mathbf{x}$; matrices, $\mathbf{X}$. This notation will be used throughout, except in specific examples where only capital characters are used.

```
 11234567890+×QWERTYUIOP←ASDFGHJKL[]ZXCVBNM,./
¨¯<≤=≥>≠∨∧-÷?⍵∊⍴~↑↓⍳○⋆→⍺⌈⌊ ∇∆∘'⎕()⊂⊃∩∪⊥⊤|;:\
```

FIGURE 1. A typical complete character set appearing on a computer print-out. In this paper, statements in APL will use the same notation but with the italic capitals replaced by ordinary capitals.

Arithmetic Operations

The following symbols and their arithmetic operators are used in APL: + (addition), × (multiplication), ÷ (division), − (subtraction), * (exponentiation). It should be noted that, with the exception of exponentiation, these symbols are identical to those usually defined. They differ from FORTRAN notation in that ÷ is used for division, × for multiplication and * for exponentiation.

Specification and Display

In FORTRAN the = symbol is used to indicate the operation that the variable on the left is replaced by the result found by carrying out the operations on the right. Output operations in FORTRAN may be specified by a WRITE statement followed by the logical device number and a format specification number. Unless output is to be binary, the format must be included in FORTRAN. In APL the symbol equivalent to the = is the left-pointing arrow (←), and output is indicated by the symbol □. Using Iverson's notation we have the following:

Specify ←
Display □

Although the display notation is extremely simple relative to FORTRAN, it does not permit the versatility in output that some computer users might desire. It is, however, relatively simple to arrange output in an acceptable manner. Display is assumed if no variable name is specified.

Reduction Operations

The operations indicated above, where we contrast arithmetic summation and product operations as they would be used in FORTRAN and APL, are referred to as reduction operations. Any binary operator or relation may be extended over vectors and matrices. This is a very useful and powerful extension not easily handled in other computing languages. Again, using different type to denote scalars, vectors and matrices, we define the reduction operations as follows:

(*a*) Reduction over a vector:

$$\mathbf{y} \leftarrow \circledcirc / \mathbf{x}$$

$$\mathbf{y} \equiv x_1 \circledcirc x_2 \circledcirc x_3 \circledcirc x_4 \ldots \circledcirc x_n$$

(*b*) Row reduction on a matrix **X**:

$$\mathbf{y} \leftarrow \circledcirc / \mathbf{X}$$

$$y_j \equiv x_{j1} \circledcirc x_{j2} \circledcirc x_{j3} \circledcirc x_{j4} \ldots \circledcirc x_{jn}$$

(*c*) Column reduction on matrix **X**:

$$\mathbf{y} \leftarrow \circledcirc / [1]\mathbf{X}$$

$$y_k \equiv x_{1k} \circledcirc x_{2k} \circledcirc x_{3k} \circledcirc x_{4k} \ldots, \circledcirc x_{nk}$$

(*d*) Matrix product:

$$\mathbf{C} \leftarrow \mathbf{A} \circledcirc_1 . \circledcirc_2 \mathbf{B}$$

$$c_{jk} \equiv \circledcirc_1 / \mathbf{a}_{j.} \circledcirc_2 \mathbf{b}_{.k} \quad \text{(assumed over all } j \text{ and } k\text{)}$$

In the definitions given above, the symbol ◎ represents any binary operator or relation. A few examples will serve to illustrate the advantages of the notation given above. (No distinction is made between scalars, vectors and matrices in our examples since the distinction will be clear in the accompanying notes.)

Notes	*APL Statement*
Sum a vector of elements in **A** and specify as **X**	X←+/A
Find the product of the elements in vector **A** and specify as **H**	H←×/A
Matrix multiplication of **A** times **B** and specify as **C**	C←A+.×B
Generate a 1 when successive elements of **A**, beginning at the right, are equal	U←=/A

The statement defining a complete matrix multiplication should clearly illustrate the advantage of APL compared with a language such as FORTRAN. Furthermore, the matrix multiplication statement is not difficult to understand since it indicates that first a multiplication will be made between the elements of a row in **A** and the corresponding elements in a column of **B**. The result is then summed, and the whole operation repeated over all rows and columns. In the context of the matrix product, a useful variant is the null operator defined by the symbol ∘. This symbol indicates that no operation is defined at a particular point in the matrix product. For example, if **x** and **y** are two column vectors of equal order, the product **xy**′ would be specified as:

$$\mathbf{x} \circ . \times \mathbf{y}$$

The result would be a matrix of order $n \times n$ containing the n^2 cross-product terms.

Concatenation

Very frequently it is required that a vector or a set of scalars be 'joined' to form new vectors. In this instance the comma will cause variables to be concatenated as illustrated below:

$$\mathbf{x} \leftarrow \mathbf{u}, \mathbf{v}, a$$

$$\mathbf{x} \equiv u_1, u_2, u_3, \ldots, u_{n_u}, v_1, v_2, \ldots, v_{n_v}, a$$

In this example the variable a has been concatenated to the vector **v** and the result concatenated to the vector **u**. The resulting vector of $n_u + n_v + 1$ elements is specified as **x**.

Rank and Restructuring Operations

It is often necessary to determine the number of elements in a vector or the order of a matrix, or to rearrange the elements of a vector to form a matrix. For such operations a monadic (single argument) rank operator and a dyadic (two argument) restructuring operator have been defined.

(*a*) *Monadic Rank Operator*

The single argument rank operator is defined in the following manner:

If $$\mathbf{x} = x_1\ x_2\ x_3\ x_4\ \ldots\ x_n$$

then n, the order of the vector, is given by

$$\rho\mathbf{x},$$

and if $$\mathbf{X} = \mathbf{x}_1\ \mathbf{x}_2\ \mathbf{x}_3\ \mathbf{x}_4\ \ldots\ \mathbf{x}_c$$

then r,c is given by

$$\rho\mathbf{X}$$

As an example of the use of the rank operator, the arithmetic mean would be stated in APL notation as illustrated in the following example.

Notes	*APL Statement*
Specify **X** as the vector of data	X←1,2,3,4,5
Divide all elements of **X** by the number of observations; sum the result	MEAN←+/X÷ρX
Display the mean	□←MEAN

It should be evident that in this APL interpreter no DIMENSION statements are required, and that DO loops do not appear where they would not usually be expected.

(*b*) *Dyadic Restructuring Operator*

A vector of elements may easily be replicated or restructured by using the dyadic ρ operator. The left-hand argument may contain either a scalar to indicate the number of times the right-hand argument is to be replicated, or two scalars indicating that the right-hand argument is to be restructured to form a matrix of the order defined by the left-hand argument.

Notes	*Result*	*APL Statement*
Specify **X** as 5		X←5
Specify **Y** as the vector **1, 2, 3**		Y←1,2,3
Display **X** replicated twice	5 5	□←2ρX
Display **Y** restructured as three rows and and three columns	1 2 3 1 2 3 1 2 3	□←(3,3)ρY
Display as three rows and three columns the vector **1, 0, 0, 0**	1 0 0 0 1 0 0 0 1	□←(3,3)ρ1,0,0,0

S.P.

In the last example, restructuring the vector proceeds in such a manner that elements not forming one line are used to form the next.

Interval Vector Generation and Index Operations

It is very useful to be able to generate a vector such that

$$\mathbf{x} = 1, 2, 3, \ldots, n.$$

This vector would be useful for generating subscripts, operations involving factorials, although in APL a factorial operator has been defined, and where a function requires evaluation between some sets of limits. The ι operator is used for this purpose. We may define this operation as follows:

$$\mathbf{x} \leftarrow \iota n$$

$$\mathbf{x} \equiv 1, 2, 3, \ldots, n.$$

Notes	*Result*	*APL Statement*
Specify the vector **X** as 1 2 3 4 5		$X \leftarrow \iota 5$
Specify a vector as 1 2 3 4 5 to which minus 3 has been added	$^-2\ ^-1\ 0\ 1\ 2$	$\square \leftarrow {}^-3 + \iota 5$

If a student wished to evaluate a function such as x^2 for values between minus 2 and plus 2 he might proceed in the following manner:

$$({}^-3 + \iota 5) * 2$$

If he wished to find 50! the solution is as simple as typing

$$\times / \iota 50$$

The definition of the dyadic interval operator is a little more complex. If we define a vector **y** and a vector **x**, which need not necessarily be of the same order, then we might define the statement

$$\mathbf{x} \iota \mathbf{y}$$

The result would be a vector of the order of **y** containing the subscript position in **x** for which each element in **y** is equal to an element in **x**. If an element in **y** is not contained within **x**, the subscript position for that element is defined as $(\rho\mathbf{x})+1$. For example, in the following operation

$$\square \leftarrow (1,3,4,2) \iota 3,4$$

the result displayed would be a vector of two elements containing the scalars 2 and 3. The 3 of the right vector is found in the second position of the left vector, and the 4 of the right vector is found in the third position of the left vector.

Vector and Matrix Subscripts

In all the examples to this point, vectors and matrices have been used without reference to subscripts. When required, subscripts for either vectors or matrices may be defined by enclosing them in square brackets. The delimiter between row and column subscripts is the semicolon. The subscript notation permits referencing a complete row or column of a matrix. In addition, APL permits subscripts to be vectors.

Notes	*Result*	*APL Statement*
Specify a matrix **M**		M←(3,3)ρ1 2 3 4 5 6 7 8 9
Display element m_{12}	2	M[1;2]
Display the second column	2 5 8	M[;2]
Display elements m_{13} and m_{23}	3 6	M[1,2;3]

Floor and Ceiling Operations

The floor operator $\lfloor$ may be defined as follows:

$$k \leftarrow \lfloor x$$

such that

$$k \leqslant x < k+1.$$

If the vector **x** replaces the scalar x the result **k** is a vector containing the floor elements of **x**. If, however, we use

$$k \leftarrow \lfloor/\mathbf{x}$$

then the smallest element of **x** is specified as k.

The ceiling operator can be defined in a similar manner. In the case of the ceiling operator $\lceil$,

$$k \leftarrow \lceil x$$

will result in k being defined as

$$k-1 < x \leqslant k.$$

If we use the vector **x**, then

$$k \leftarrow \lceil/\mathbf{x}$$

will specify k as the largest element of the vector **x**.

Notes	*Result*	*APL Statement*
Specify a vector **X**		X←2.6,3.1,6.2
Display the floor of **X**	2 3 6	⌊X
Display the minimum element of **X**	2.6	⌊/X
Display the ceiling of **X**	3 4 7	⌈X
Display the maximum element of **X**	6.2	⌈/X

Compression Operation

Very frequently it is required to remove certain elements from a vector and to 'compress' the remaining elements. This operation removes elements and changes the order of the vector. If a binary vector **u** is defined as containing a 0 in the index position corresponding to the position from which an element is to be removed from a second vector, and 1's in the index position where elements are to be retained, we may define the compression operation as follows:

$$\mathbf{z} \leftarrow \mathbf{u}/\mathbf{x}$$

and the result will produce a vector **z** containing only those elements of **x** which were not suppressed by the 0's in **u**.

Notes	*Result*	*APL Statement*
Specify the vector **X**		X←2,8,4,3,9
Suppress even elements of the vector **X**	2 4 9	(1,0,1,0,1)/X

Factorial Operation

The factorial operator is identical in APL notation to that usually used in mathematics. Its use in the system requires that the operator appear to the left of the scalar (or vector) on which it operates. For example,

FACT←(!50)÷!25

will specify FACT as the value of 50! divided by 25!. Since this is a very large number, our particular interpreter displays the result in exponential format.

Relation Operations

All of the relations < ⩽ = ⩾ > and ≠ in addition to the logical operators ∨ and ∧ are available. Results produced are binary, i.e. a 1 if the condition is true and a 0 if the condition is false. These operators are very important for determining the conditions of branching and for certain tabulation operations. Our example below shows how a simple frequency tabulation may be carried out. The use of these operators in branching statements is reserved for the next section.

Notes	*Result*	*APL Statement*
Specify vector		X←3,2,4,6,2,2,1,4,6,5
Tabulate	1 3 1 2 1 2	FREQ←+/(ι6)∘.=X

FREQ, in the above example, will contain the correct frequency tabulation of the elements of the vector **x**. The right-hand portion of the statement, (ι6)∘.=X, will generate the following matrix of 1's and 0's:

$$\begin{array}{cccccccccccc}
 & & & & & & \mathbf{X} & & & & & \\
 & & 3 & 2 & 4 & 6 & 2 & 2 & 1 & 4 & 6 & 5 \\
 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
 & 2 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\
 & 3 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\iota 6 & 4 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
 & 5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
 & 6 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0
\end{array}$$

The final operation, a reduction operation (+/) over the rows of the matrix, sums the rows to produce the frequencies 1, 3, 1, 2, 1, 2. It is a very simple matter to determine the maximum and minimum values and then to generate the appropriate row categorization when the numbers to be tabulated do not form a continuous sequence beginning at one. It should be noticed that the right-hand statement producing the matrix to be tabulated by row, is of the general matrix product form. In this sense, Iverson's notation is markedly superior to existing mathematical notation.

Function Definition and Branching

Since APL requires an interpreter and executes each APL statement immediately, the statement itself is lost unless special instructions are given indicating that a set of statements is to be saved. Therefore branch statements can only be used within a set of APL statements which have been specifically labelled as requiring complete retention even after execution. Readers will be familiar with this concept since it is essentially the same as functions in other programming languages.

The specification of a function is made in the following manner:

$$\mathrm{R} \leftarrow \mathrm{A1\ NAME\ A2}$$

In this instance the function is called by the name NAME and takes in two arguments specified as A1 and A2. A1 and A2 may be scalars, vectors or matrices. The result is returned in R. The sumbol ▽ must appear as the first and last symbol to specify that a function is being defined.

Branching statements can only be made within functions since without a definition of a function, a statement is lost upon execution. The general form of a branch statement is as follows:

$$\rightarrow (x\ \mathrm{R}\ y)/s$$

This is interpreted as 'go to the *s*th statement if the relation specified by R between *x* and *y* is true, otherwise go to the next statement'. For example, in the statement

$$\rightarrow (5 = \mathrm{CTR})/7$$

if CTR was equal to 5 the seventh statement would be executed, otherwise the next statement would be executed. Arithmetic expressions are permitted within the parentheses. If, in the following example, **x** is a vector, then

$$\rightarrow (1 \leqslant +/0 = x)/4$$

would cause the fourth statement to be executed if **x** contained one or more zeros. Symbolic addresses may also be used in place of the numeric ones used in our examples.

3. APL PROGRAMMING EXAMPLES

Example 1: *Sign Test*

```
     ∇PROB←X ANSIG Y
[1]  PND←⌊/(+/0<X−Y),+/0>X−Y
[2]  N←(ρX)−+/0=X−Y
[3]  R←(ιPND+1)−1
[4]  PROB←+/((!N)÷(!R)×!(N−R))×(0.5*R)×0.5*N−R
[5]  PND,N
[6]  ∇
```

The following explanation is given for each statement number:

[1] Find the number of positive and negative differences and specify PND as the smaller of the two.
[2] Specify N as the number of elements in **X** less the number of zero differences between **X** and **Y**.
[3] Specify **R** as a vector of event categories as used in the binomial equation $(N!/\mathbf{R}!(N-\mathbf{R})!)P^{R}Q^{N-R}$
[4] Calculate the probabilities and sum.
[5] Display PND and N. PROB is displayed upon the function being completed.

Example 2: *Kruskal–Wallis Analysis of Variance*

```
     ∇H←NG KWA X
[1]  RO←RANK X
[2]  G←ι0
[3]  CON←0
[4]  K←1
[5]  G←G,+/RO[CON+(ι(NG[K]))]
[6]  →(K=ρNG)/10
[7]  CON←NG[K]+CON
[8]  K←K+1
[9]  →5
[10] H←((12÷((ρRO)*2)+ρRO)×(+/(G*2)÷NG))−3×(ρRO)+1
[11] ∇

     ∇R←RANK X
[1]  XRO←X[S←(+/(X∘.>X)+(X∘.=X)∧((ιρX)∘.≤ιρX))ιιρX]
[2]  TR←((ιρX)+.×(XRO∘.=XRO))÷+/XRO∘.=XRO
[3]  R←TR[SιιρX]
[4]  ∇
```

The calling statement of the program is specified in the statement H←**NG** *KWA* **X**. The value H is returned by the program while the entering arguments are **NG** and **X**. For this function **X** is a vector containing the observed measures of all observations in order of the groups to which they belong. **NG** is a vector containing the number of observations in each group, ordered according to the data for each group in **X**. The number of elements in **NG**, ρNG, indicates the number of groups that are contained within the data defined by the vector **X**, while ρX is equal to the total sample size. The function KWA calls upon the function RANK to assign rank values to the vector of observations **X**. When this is completed the vector is returned as RO and KWA then proceeds to find the appropriate summed ranks for each group.

Example 3: *Correlations*

```
      ∇CORREL
 [1]  'THIS FUNCTION WILL ACCEPT BY ROW AN OBSERVATION BY
      VARIABLE SCORE MATRIX'
 [2]  'PLEASE TYPE THE NUMBER OF VARIABLES NEXT'
 [3]  'NV←□
 [4]  R←(NV,NV)ρ0
 [5]  'NOW TYPE THE NUMBER OF OBSERVATIONS'
 [6]  N←□
 [7]  'NOW TYPE THE VARIABLES FOR EACH OBSERVATION.  WHEN
      FINISHED TYPE ¯99.'
 [8]  M←NVρ0
 [9]  S←NVρ0
[10]  →12
[11]  'ERROR RETYPE LAST LINE - - - - - LAST LINE IGNORED'
[12]  X←□
[13]  →(1=+/¯99=X)/END
[14]  →(NV≠ρ,X)/11
[15]  R←R+X∘.×X
[16]  S←S+X*2
[17]  M←M+X
[18]  →12
[19]  END :  'MEANS'
[20]  ''
[21]  □←M←M÷N
[22]  'STANDARD DEVIATIONS'
[23]  ''
[24]  S←((S÷N)−M*2)*0.5
[25]  S
[26]  'CORRELATIONS'
[27]  ''
[28]  R←((R÷N)−M∘.×M)÷S∘.×S
[29]  R←(⌊100×R+0.005)÷100
[30]  ∇
```

Titles which are to be displayed are enclosed in single quotation marks. Data to be accepted while on-line are indicated by a variable name being specified as □. Statement number 13 checks to determine whether −99 has been input, and if it has, termination of data input is made. Statement 14 checks to determine whether the correct number of variables have been input. If an incorrect number is sensed, control is transferred to an error message and a request is made to retype the last line. Statement 15 calculates the matrix of sums of squares and cross-products and accumulates the results over all observations in the matrix **R**. Statement 28 is the matrix equivalent of

$$r_{jk}=\{(\Sigma x_j x_k/N)-\bar{x}_j\bar{x}_k\}/\sqrt{(s_j s_k)}.$$

Statement 29 rounds-off the correlation coefficients to two digits.

Example 4: *Computer Assisted Instruction*

The power and versatility of APL is not limited to numerical calculations, as the example below will illustrate. Although the use of the computer as a direct teaching aid is not new, tremendous problems exist in retraining subject-matter specialists to program for the computer at least some portion of their course work. APL appears sufficiently easy to learn, relative to such languages as the IBM Coursewriter II, that the initial success so important to having a course author prepare material for computer instruction, is easily obtained. The example given below is intended to show how alphabetic information can be handled in the APL system. The final result of the program is similar to the result provided by the editing function of Coursewriter II.

```
    ∇FC←CHECK ANS
[1] ANSC←'ANAPLASTIC CARCINOMA'
[2] ANSR←'CARCINOMA ANAPLASTIC'
[3] CTRC←+/ANSC=ANS
[4] CTRR←+/ANSR=ANS
[5] →(CTRC>CTRR)/8
[6] FC←'–CARCINOMA ANAPLASTIC'[1+(ιρANSR)×ANSR=ANS]
[7] →0
[8] FC←'–ANAPLASTIC CARCINOMA'[1+(ιρANSC)×ANSC=ANS]
[9] ∇
```

If the student responds with the words ANAPLASTIC CARCINOMA or CARCINOMA ANAPLASTIC and one of these answers is conveyed to the checking function CHECK, the correct answer is displayed. If, however, the student responded with, say, ANEPLISTIC DARCINOME the function would return AN–PL–STIC –ARCINOM–. The power of APL is of course embodied in statements 6 and 7 in the above example. Characters enclosed between single quotation marks are stored by the system as alphanumeric characters. It should be noticed that alphanumeric vectors may be subscripted.

4. CONCLUSION

There is no question in the mind of the writer that APL, as devised by Iverson, is a language that should most certainly receive considerable attention of not only researchers with specialized mathematical problems, but also by teachers at all levels of instruction. The main advantages of APL rest with its precise and meaningful symbolism, and its applicability in the classroom as well as with the computer. There would also appear to be merit in the use of APL for purposes of mathematics instruction and the construction of computer-assisted teaching programs.

REFERENCES

IVERSON, K. E. (1962). *A Programming Language.* New York : Wiley.
IVERSON, K. E. (1966). *Elementary Functions.* Chicago: Science Research Associates.

COMPUTER PROGRAM ABSTRACTS

Text Sorting Programs. By J. R. Birnie, Department of Education, University of Durham.

Description *UA2 Sorting Big* A program designed to scan text and abstract details of:

1. sentence length,
2. different words used in each sentence,
3. numbers of linkers operating at clause level and above (linker name output) in each sentence,
4. numbers of free linkers (linker name output) in each sentence,
5. (optional) errors in sentence,
6. cumulative details of 1, 2, 5 at the end of each text in terms of errored and error-free text,
7. ordered word-frequency list.

The program will handle texts containing up to 2,000 different orthographical words. A dictionary of linkers is read in as data tape 1. Errors must be numbered by type in the text at the data-preparation stage, otherwise it will be assumed that all text is error-free.

UA4 Sorting As above but facilities 1, 2, 7 and cumulative details of 1 and 2 only available for *n* number of texts.

The program will handle texts containing up to 500 different words in batches of four (more if they contain fewer words).

Computers, Language Both programs are available in ALGOL suitable for running on the KDF9. A version exists of UA2 suitable for running on the Elliott 803B but is not recommended as output (paper tape) is slow.

Data Format, Output Input is eight-hole paper tape, output on the KDF9 is from a line-printer with a pre-set format. Words beginning with capitals are ignored, therefore lower-case characters must be used at the beginning of sentences. A title for each text is allowed and this title will be output at the head of the appropriate set of results. It must be enclosed by exponential signs. Each text must be terminated with a multiplication sign and the final character read must be the normal end message character.

Computation Times UA2 deals with approximately 100 words per minute; UA4 about five times as many.

Availability Program tapes and data tape, together with data preparation notes, are available from the Research Secretary, Department of Education, University of Durham.

Reference

Bateni, M. R. (1959). Linkage. (Unpublished M.A. thesis, University of Leeds.)

Slater's *i* [and nearest adjoining orders] from Paired Comparisons or Balanced Incomplete Block Design Rankings. By J. P. N. Phillips, Department of Psychology, University of Hull.

Description Program A, the full program, takes as data a set of paired comparisons or, following Schucker (1959) and Gulliksen and Tucker (1961), a set of preference rankings (arranged in a balanced incomplete block design) [and a list of object descriptions]. It checks the data for obvious errors: if any are found, a diagnosis of each is output and no

Example 4: *Computer Assisted Instruction*

The power and versatility of APL is not limited to numerical calculations, as the example below will illustrate. Although the use of the computer as a direct teaching aid is not new, tremendous problems exist in retraining subject-matter specialists to program for the computer at least some portion of their course work. APL appears sufficiently easy to learn, relative to such languages as the IBM Coursewriter II, that the initial success so important to having a course author prepare material for computer instruction, is easily obtained. The example given below is intended to show how alphabetic information can be handled in the APL system. The final result of the program is similar to the result provided by the editing function of Coursewriter II.

```
    ∇FC←CHECK ANS
[1]  ANSC←'ANAPLASTIC CARCINOMA'
[2]  ANSR←'CARCINOMA ANAPLASTIC'
[3]  CTRC←+/ANSC=ANS
[4]  CTRR←+/ANSR=ANS
[5]  →(CTRC>CTRR)/8
[6]  FC←'–CARCINOMA ANAPLASTIC'[1+(ιρANSR)×ANSR=ANS]
[7]  →0
[8]  FC←'–ANAPLASTIC CARCINOMA'[1+(ιρANSC)×ANSC=ANS]
[9]  ∇
```

If the student responds with the words ANAPLASTIC CARCINOMA or CARCINOMA ANAPLASTIC and one of these answers is conveyed to the checking function CHECK, the correct answer is displayed. If, however, the student responded with, say, ANEPLISTIC DARCINOME the function would return AN–PL–STIC –ARCINOM–. The power of APL is of course embodied in statements 6 and 7 in the above example. Characters enclosed between single quotation marks are stored by the system as alphanumeric characters. It should be noticed that alphanumeric vectors may be subscripted.

4. Conclusion

There is no question in the mind of the writer that APL, as devised by Iverson, is a language that should most certainly receive considerable attention of not only researchers with specialized mathematical problems, but also by teachers at all levels of instruction. The main advantages of APL rest with its precise and meaningful symbolism, and its applicability in the classroom as well as with the computer. There would also appear to be merit in the use of APL for purposes of mathematics instruction and the construction of computer-assisted teaching programs.

References

Iverson, K. E. (1962). *A Programming Language*. New York : Wiley.
Iverson, K. E. (1966). *Elementary Functions*. Chicago: Science Research Associates.

COMPUTER PROGRAM ABSTRACTS

Text Sorting Programs. By J. R. Birnie, Department of Education, University of Durham.

Description *UA2 Sorting Big* A program designed to scan text and abstract details of:

1. sentence length,
2. different words used in each sentence,
3. numbers of linkers operating at clause level and above (linker name output) in each sentence,
4. numbers of free linkers (linker name output) in each sentence,
5. (optional) errors in sentence,
6. cumulative details of 1, 2, 5 at the end of each text in terms of errored and error-free text,
7. ordered word-frequency list.

The program will handle texts containing up to 2,000 different orthographical words. A dictionary of linkers is read in as data tape 1. Errors must be numbered by type in the text at the data-preparation stage, otherwise it will be assumed that all text is error-free.

UA4 Sorting As above but facilities 1, 2, 7 and cumulative details of 1 and 2 only available for *n* number of texts.

The program will handle texts containing up to 500 different words in batches of four (more if they contain fewer words).

Computers, Language Both programs are available in ALGOL suitable for running on the KDF9. A version exists of UA2 suitable for running on the Elliott 803B but is not recommended as output (paper tape) is slow.

Data Format, Output Input is eight-hole paper tape, output on the KDF9 is from a line-printer with a pre-set format. Words beginning with capitals are ignored, therefore lower-case characters must be used at the beginning of sentences. A title for each text is allowed and this title will be output at the head of the appropriate set of results. It must be enclosed by exponential signs. Each text must be terminated with a multiplication sign and the final character read must be the normal end message character.

Computation Times UA2 deals with approximately 100 words per minute; UA4 about five times as many.

Availability Program tapes and data tape, together with data preparation notes, are available from the Research Secretary, Department of Education, University of Durham.

Reference

Bateni, M. R. (1959). Linkage. (Unpublished M.A. thesis, University of Leeds.)

Slater's *i* [and nearest adjoining orders] from Paired Comparisons or Balanced Incomplete Block Design Rankings. By J. P. N. Phillips, Department of Psychology, University of Hull.

Description Program A, the full program, takes as data a set of paired comparisons or, following Schucker (1959) and Gulliksen and Tucker (1961), a set of preference rankings (arranged in a balanced incomplete block design) [and a list of object descriptions]. It checks the data for obvious errors: if any are found, a diagnosis of each is output and no further analysis attempted. Otherwise it determines and outputs *i*, the minimum number of inconsistencies in the data [and the object descriptions in all nearest adjoining orders obtained by rectifying the inconsistencies] (Slater, 1961), using a procedure which is most efficient with relatively consistent data (Phillips, 1967). The shorter program B does not carry out the operations listed in square brackets.

Language Elliot 803 ALGOL.

Data Format The data layout required is an almost exact copy of a convenient form of experimental protocol.

Space Required and Computation Time Both are effectively unpredictable, since the program uses a recursive procedure, which is re-entered an unpredictable number of times, each time setting up arrays of unpredictable dimensions: however, both will be negatively correlated with the consistency of the data.

Availability Any of the following documents for each program may be obtained from the author:

(i) The print-out of an ALGOL *comment*, giving full instructions for use of the program, with examples.
(ii) The print-out of the whole program, including (i).
(iii) The Elliot 803 five-hole tape of (ii).
(iv) The Elliot 803 fast check binary tape of the program.

For users with access to an Elliot 803 with 8,192 word-store, (i) and (iv) will probably be most convenient. For those with access to an installation which accepts Elliot 803 ALGOL in 5-hole tape, (iii) will be applicable. For all other users, (ii) is most convenient.

References

GULLIKSEN, H. and TUCKER, L. R. (1961). A general procedure for obtaining paired comparisons from multiple rank orders. *Psychometrika* **26**, 173–183.

PHILLIPS, J. P. N. (1967). A procedure for determining Slater's *i* and all nearest adjoining orders. *Br. J. math. statist. Psychol.* **20**, 217–225.

SCHUCKER, R. E. (1959). A note on the use of triads for paired comparisons. *Psychometrika* **24**, 273–276.

SLATER, P. (1961). Inconsistencies in a schedule of paired comparisons. *Biometrika* **48**, 303–312.

BOOK REVIEWS

Advanced Theory of Statistics. Vol. III. By M. G. KENDALL and A. STUART. London: Charles Griffin & Co. Pp. 552. 136*s*.

This is the last volume of the revision of M. G. Kendall's two-volume *Advanced Theory of Statistics* (1947–48). The original was initially planned in synopsis before the late war by a group of leading statisticians, but circumstances left it to Dr. Kendall to compile. This was fortunate in that the outcome was a book with the unity of style given by single authorship yet with the encyclopaedic compass which has established it as the standard reference book of statistical theory. During the following two decades, there has been an immense outpouring of papers which have greatly widened and deepened our knowledge of statistical theory as it was in 1947 without substantially adding any new major statistical ideas. There have been, it is true, various all-embracing schemes: Decision theory, Information theory, Games theory, Neo-Bayesian theory, etc., which, while throwing new light on existing practice, have expired or are expiring without substantially altering that practice. Perhaps Sequential Analysis, with its peculiar aptness in clinical trials, is the only new idea contributed to statistical methodology which has found a permanent niche. But it was really a war-time development. This third volume completes the formidable task which Dr. Kendall took on, with Professor Stuart in partnership, of building the *statistical* part of this huge corpus of papers into the original scheme.

The book itself is, in essence, four separate books in one binding. That is, it consists of four distinct sections, each substantially self-contained, although the first leans somewhat on the theory of least squares developed in Volume II. They are: Analysis of Variance (pp. 1–165), Survey Theory (pp. 166–238), Multivariate Theory (pp. 239–365) and Time Series Analysis (pp. 366–503). On stylistic grounds one may guess that the first two sections owe more to Professor Stuart and that the last two are largely Dr. Kendall's. The former are distinguished by a fairly exhaustive coverage of the literature and tend as a consequence to have many of the properties, both good and less good, of a review. The great advantage of this form is particularly seen in the first section where, after giving a more rounded analytical general theory than was possible in 1948 and treating fully the components-of-variance model which now has a fairly well developed theory, space is found for the ragged edges of the Analysis of Variance where the present state of knowledge is inconclusive or altogether less cut and dried. This miscellany of topics includes notably: multiple comparisons and questions of design; questions of relaxation of the standard assumptions and techniques of transformation of variables intended to make these hold more nearly.

The section on survey theory is an excellent compression of the often intricate though elementary sampling theory in this field. It also illustrates one drawback of the review approach, namely the very occasional lapse into uncritical reporting. For example, a page is devoted to a précis of Pathak's nebulous ' definition ' of Sufficiency in finite population sampling. This is an attempt to generalize a valuable insight, and the application given here illustrates this insight but does not test the wider aspects of the general definition, so that the example tells us more than the (redundant) ' definition '.

Both these two sections give a good guide to the present state of the theory, which is less true of the latter two whose virtue is their more organic unity. This is doubtless partly due to the fact that they are direct expansions of the corresponding sections in Dr. Kendall's original book and partly perhaps to the more unitary nature of their subject matter, but the approach in these sections is noticeably more selective.

The section on multivariate analysis is possibly the best concise development of the basic distribution theory that exists. It is very strongly to be recommended to readers new to the field who want to master the key mathematical results before tackling the more complex statistical issues involved. The treatment of discriminant analysis is not really updated, although some of the relevant papers are cited, and there is a notable omission of the canonical discrimination methods of Healy and of Bartlett and his students. The discussion here is misleading and, surprisingly, the common statistical trouble with discriminant analysis (namely that the individual to be classified belongs to none of the populations in question) is not included in the discussion of contingencies.

The time series analysis here, as in the original Volume II, has the unique distinction of giving an adequate exposition of the mathematical theory yet remaining fundamentally statistical, which is to say relevant to practice and applicable. It is a pleasant change from the usual dreary exercise in formal symbol pushing in pursuit of unrealistic models. Some of the rather technical moving-average methods beloved of old-school economists still persist in this volume and could have been further tidied up and tied in with modern theory. Also, a more careful distinction between statistic and parameter, in places, would have made the argument clearer. And one would have supposed that Laplace's *Théorie Analytique* would have been sufficiently familiar to the authors to dispel their belief that saddle-point methods were first systematically introduced into statistics 144 years after that book was written. But all these are negligible criticisms when compared with the major virtue of this section.

The whole of this volume continues the refreshingly civilized approach of the two preceding ones—it is fully motivated and practically relevant; it is not afraid of heavy mathematics where necessary, but keeps it in its place and avoids any for its own sake. The price, in pounds sterling, is considerable but the book is much better value than most. No print is wasted on Carnaby Street statistics.

D. E. BARTON.

Elements of Non Parametric Statistics. By GOTTFRIED E. NOETHER. New York and London: John Wiley & Sons Inc. 1967. Pp. x + 104. 64*s*.

Since the book only totals some one hundred pages, Professor Noether is inevitably faced with the acute problem of the selection of topics for inclusion. Nevertheless, by adopting a unified treatment of his material wherever possible and using mostly only sketches of proofs, he contrives to cover a number of the basic features of this fast-expanding field.

A large part of the book is concerned with an integrated development of the theory of the Rank Sum Test. This allows the author to cover one- and two-sample location problems, the two-sample scale problem, and to treat more complex problems such as the analysis of balanced incomplete block designs using integer ranks. The simpler problems are considered both from the point of view of tests and from the wider concept of confidence intervals for the parameters concerned.

Some space is devoted to the uses of order statistics, including the generation of confidence and tolerance intervals, and there is a brief but quite elegant treatment of some of the better known distribution-free tests of fit. The book is completed by a chapter on the measure of association known as Kendall's Tau, and some discussion of the efficiency of non-parametric methods using the local asymptotic measure devised by Pitman.

The author's stated objective is not to produce a text which is a 'cook-book' of apparently disconnected non-parametric methods but rather to develop sufficient of the theory to help the reader with some mathematical background to gain a better understanding of the basic distribution-free techniques. I feel that he has achieved his aim but could have done so more convincingly if he had covered the field more extensively. This, of course, would have meant considerably more space. Even in a book of the present length, however, I would have liked to have seen some mention of the use of scores other than integer ranks, and more comments on the somewhat embarrassing multiplicity of non-parametric methods that are available for any of the simple problems discussed. It is a pity also that no mention is made of permutation tests using the sample values themselves since, in certain circumstances, these techniques can throw considerable light on the 'normal-theory' methods.

Professor Noether has a pleasant style of presentation and the mathematical level will certainly not deter any of his intended audience. The book certainly contains material and emphases not to be found in other texts on non-parametric methods. For instance, he includes a quite detailed examination of the standard methods when the underlying random variable is discrete; and of the closely connected problem of ties. Also, the degree of emphasis placed on distribution-free interval estimation is unusual and pleasing; and the same remark applies to his integrated development of the Rank Sum Test.

In conclusion, I would express the hope that following this pleasant aperitif, Professor Noether will see fit to write a much extended textbook on his subject directed to the same audience. His efforts would almost certainly be appreciated by both teachers and students of statistics.

J. G. Fryer.

Principles of Psychological Measurement. By G. C. Helmstadter. London: Methuen. 1966. Pp. xx+248. 36*s*.

This is an attempt to present, in a reasonably compact form, the present position in the theory related to this rapidly developing field. Some mathematical derivations of an elementary character are given, but, on the whole, the reader requires only a knowledge of descriptive statistical techniques and some familiarity with conventional test materials. It includes nine chapters of theory and a final overview.

The concepts of standardization, objectivity, reliability and validity have been presented within a unifying framework of error types; for example, standardization under 'interpretative' errors, objectivity under 'personal' errors, reliability under 'variable' errors, and validity under 'constant' errors. Standardization and reliability are dealt with in single chapters, but there are separate chapters on content, empirical and construct validity.

Helmstadter makes the case that the purposes of testing can all be subsumed under 'prognosis', and argues that, for example, "diagnosis is an elaborate form of prediction involving the anticipated consequences of suggested therapies or other contemplated actions". His concentration on validation techniques is consistent with this view, and in keeping with present educational trends increasingly concerned with guidance and evaluation of curricula. An early introduction of expectancy tables adds clarity to the statistical aspects of prediction.

Unfortunately, his explanation of some of the more complex concepts lacks crispness. For example, the different kinds of errors involved in test-retest and parallel-form reliability determinations are neither well defined nor well illustrated. The chapter on 'Construct Validity' uses the highly philosophical language of Cronbach and Meehl's original paper on the subject before giving an example; by contrast, Cronbach's own elementary textbook adopts a heuristic approach via the construct of anxiety before attempting a definition, which is more suited to the average student.

Chapters on Test Development and Scales and Inventories, incorporating a good deal of recent work, followed by a useful summary with diagrammatic illustrations of methods available for prediction of multiple criteria in the penultimate chapter on Multiple Measurement, may redeem the book, particularly in the eyes of students training for guidance work. It should also stimulate those readers capable of becoming specialists in the measurement field to continue their studies by reading the more detailed treatments of, for example, Gulliksen and Torgerson and the many other references collected in the five-page bibliography.

D. M. Penfold.

A Study of a Biased Friendship Net. By T. J. FARARO and M. H. SUNSHINE. Syracuse: Syracuse University Press. 1964. Pp. x+124. (No price given.)

Attempts to describe a population of persons in terms of ' random nets ' exemplify one of the problems of mathematical models of behaviour: that is, how can simple sets of symbols, stochastic or otherwise, sufficiently represent the complex interaction of the elements of human activity. Apart from those models which proliferate parameters to explain everything the problem has, so far, only been demonstrated. The present study is no exception but provides an interesting variant of this limited approach.

In the first place, the monograph makes a good attempt to explain what the mathematics is about. It effectively presents a sociometric procedure which is not well known, that of ' tracing ' in a set of persons, and proceeds to show its relationship to random and biased net theory. By relating bias parameters to certain simple structure statistics of behaviour between persons, a biased net tracing formula is derived and its consequences are discussed. This is followed by an empirical study of friendship carried out at a Junior High School where the children are almost entirely from lower income Negro families and where there is a high delinquency rate.

Systematic discrepancies were observed between the basic formula containing two bias parameters and the data. This appeared to be due to the assumption that the biased net contact density was constant at each step and evidence is then found that this is indeed the case. The trouble is, of course, that any attempt to define a net formula based upon a variable contact density would be difficult and would, presumably, involve the addition of further parameters to the model. Notwithstanding, further studies are made on sub-nets of the population and interesting features found. However, it is doubtful if the basic inadequacy of the model can be satisfactorily removed.

Hence the value of this study depends upon whether or not it illustrates the correct kind of mathematical model for studies of friendship or is, at least, a useful methodological approach to sociometric analysis. It may not be the former but it is certainly the latter. Although the basic assumptions underlying the biased net model, as developed by the present authors and others, may be inadequate because of the greater fluidity and variability in the actual interactions within the ' net ' which are described in a rather static manner by the model, it offers a starting point for anyone attempting an analysis of group activity. The authors may be complimented on a well constructed study, adequately introduced and described in as simple a manner as is possible with so complex and interesting a psychological situation.

A. R. PIKE.

Human Factors Evaluation in System Development. By DAVID MEISTER and GERALD F. RABIDEAU. New York: Wiley & Sons. 1965. Pp. xii+307. 75*s*.

This book describes the contributions of human factors studies to the evolution of man/machine systems, and the stages in the development of the system at which various methods become appropriate. It examines the working relationships between the human engineer and others who participate in achieving an efficient system.

The authors begin by attempting to establish a theoretical basis for human factors evaluation. They claim that such evaluations are made because they are " essential to system development whenever there are design alternatives affecting human performance to be decided between ". Their concept of the System Development Cycle is explained and the role of human factors is indicated in the functional analysis or planning of the system. Evaluation is seen as a verification of this analysis and " the major problem for the human engineer in evaluating system performance is to determine how intermediate, behavioural, and representational criteria interact with and influence terminal performance and mechanistic criteria ".

The second part of the book considers the process of design verification leading to testing the system as a whole, surveys the processes and problems in planning performance evaluations, and classifies data collection methods. The analysis of human factors data in system evaluation and the application of human factors techniques to the manufacturing processes for system components are described. The book ends with a brief survey of the problems the human engineer can expect to encounter in fulfilling his rọle in system development.

The idea behind this book is well conceived: a thorough appraisal of the topics covered was needed. Yet the book is disappointing. The preface, intended to describe what is covered and what is excluded, is earnestly and carefully written but somehow fails to make quite clear either the contents or the intentions. Opening the book at random usually reveals a series of general statements, bordering on homilies, which are often well worth making but are seldom pursued or enlightened by examples. Page after page of general statements makes uninspired, unmemorable, and finally, dull reading. Frequently thoughts already expressed are rephrased and repeated, although the style of writing makes this less obvious than it might otherwise be. When examples are included they refer to " a hypothetical system, the Hermes space tug ". The intention to achieve continuity and coherence by examples relating to a single system is not fulfilled, and examples drawn from real systems might have made the issues clear and permitted better illustrations. The photographs are largely irrelevant and very badly reproduced.

Several chapters, particularly in the second part, contain much useful material, in the form of distinctions between concepts, tabulations of factors relevant to a particular issue, classifications of criteria, and discussions of basic terms. Such material is in the main thorough, painstaking and practical. Unfortunately these merits which the book possesses tend to be overlooked because of its inadequacies. V. David Hopkin.

INDEX TO VOLUME 20

ARTICLES

Author	*Title*	*Page*
Bainbridge, L.	A Discrete Predicting Controller	31
Baker, F. B.	The Effect of Criterion Score Grouping upon Item Parameter Estimation	227
Constantinescu, P.	A Method of Cluster Analysis	93
Gregson, R. A. M. and Russell, P. N.	A Note on a Generating Assumption in McGee's Multidimensional Analysis of 'Elastic' Distances	239
Hamilton, M.	Comparison of Factors by Ahmavaara's Method	107
Hunka, S.	APL: A Computing Language Designed for the User	249
Jonckheere, A. R. and Bower, G. H.	Non-Parametric Trend Tests for Learning Data	163
Lockhart, R. S.	The Assumption of Multivariate Normality	63
McDonald, R. P.	Factor Interaction in Nonlinear Factor Analysis	205
McGee, V. E.	A Reply to Some Criticisms of Elastic Multidimensional Scaling	243
Peel, E. A.	A Method for Investigating Children's Understanding of Certain Logical Connectives used in Binary Propositional Thinking	81
Phillips, J. P. N.	A Procedure for Determining Slater's *i* and all Nearest Adjoining Orders	217
Rodger, R. S.	Type I Errors and their Decision Basis	51
Rodger, R. S.	Type II Errors and their Decision Basis	187
Shallice, T.	Temporal Summation and Absolute Brightness Thresholds	129
Stringer, P.	Cluster Analysis of Non-Verbal Judgements of Facial Expressions	71
Thomas, E. A. C.	Reaction-Time Studies: The Anticipation and Interaction of Responses	1

BOOK REVIEWS

Author	*Title*	*Page*
Anstey, E.	Psychological Tests	123
Fararo, T. J. and Sunshine, M. H.	A Study of a Biased Friendship Net	266
Frederiksen, N. and Gulliksen, H. (eds.)	Contributions to Mathematical Psychology	116
Green, D. M. and Swets, J. A.	Signal Detection Theory and Psychophysics	112
Helmstadter, G. C.	Principles of Psychological Measurement	265
Hudson, L.	Contrary Imaginations	111
Kalton, G.	Introduction to Statistical Ideas for Social Scientists	121
Kelley, H. P.	Memory Abilities: a Factor Analysis	122
Kendall, M. G. and Stuart, A.	Advanced Theory of Statistics, Vol. III	263
Kish, L.	Survey Sampling	111
Meister, D. and Rabideau, G. F.	Human Factors Evaluation in System Development	266
Noether, G. E.	Elements of Non-Parametric Statistics	264
Oppenheim, A. N.	Questionnaire Design and Attitude Measurement	122
Ross, S.	Logical Foundations of Psychological Measurement	114
Seal, H. E.	Multivariate Statistical Analysis for Biologists	120
Uhr, L. (ed.)	Pattern Recognition: Theory, Experiment, Computer Simulations and Dynamic Models of Form Perception and Discovery	115
Zetterberg, H. L.	On Theory and Verification in Sociology (3rd ed.)	121

COMPUTER PROGRAM ABSTRACTS

Author	*Title*	*Page*
Birnie, J. R.	Text Sorting Programs	261
Carlson, J. E.	A FORTRAN Program for Nonparametric Statistics	126
Hallworth, H. J. and Brebner, A.	Factor Scores Programs	127
Hunka, S. and Wahlstrom, M.	A FORTRAN Program for Canonical Correlations	125
Phillips, J. P. N.	Slater's *i* [and nearest adjoining orders] from Paired Comparisons or Balanced Incomplete Block Design Rankings	261
Wyatt, M.	Program for Factor Analysis	125

PUBLICATIONS OF THE BRITISH PSYCHOLOGICAL SOCIETY

The British Journal of Psychology (about 512 crown quarto pages per volume), *The British Journal of Medical Psychology* (about 350 crown quarto pages per volume), and the *British Journal of Social and Clinical Psychology* (about 320 crown quarto pages per volume) appear in quarterly parts, the subscription price of each, payable in advance, being £4 10*s*. per volume post-free ($16.50 in the U.S.A.). Subscriptions to any of these journals may be sent to any bookseller, or subscription agent, or to the Cambridge University Press, Bentley House, 200 Euston Road, London, N.W.1. American enquiries should be addressed to the Cambridge University Press, American Branch, 32 East 57th Street, New York, N.Y. 10022.

The British Journal of Mathematical and Statistical Psychology is issued twice a year, in May and November, in volumes of about 288 pages. The subscription rate, payable in advance , is £3 10*s*. per volume post free ($13 in the U.S.A.). Subscriptions should be sent to the Society's Agents, Taylor and Francis, Ltd., Red Lion Court, Fleet Street, E.C.4; or to any bookseller.

The British Journal of Educational Psychology is issued three times a year, in February, June and November. The subscription is £2 a year post free, payable in advance, or £1 each number. The subscription price in the U.S.A. is $4.20. Subscriptions should be sent to the Manager, B.J.E.P. Department, Methuen and Co. Ltd., 36 Essex Street, Strand, London, W.C.2.

A series of *Monograph Supplements* is issued in connexion with *The British Journal of Psychology* but is not included in the subscription for that journal. Orders may be sent to any bookseller, or the Cambridge University Press, 200 Euston Road, London, N.W.1, from whom details of these publications may be obtained.

MANUSCRIPTS

for *The British Journal of Psychology* should be sent to
Professor W. Sluckin, Department of Psychology, The University, Leicester;
for *The British Journal of Medical Psychology* should be sent to
Dr. T. Freeman, Dundee Royal Liff Hospital, By Dundee;
for *The British Journal of Social and Clinical Psychology* should be sent to
Miss Thelma Veness (Social Psychology Editor), Department of Psychology, Birkbeck College, Malet Street, London, W.C.1, *or* to
Dr. R. R. Hetherington (Clinical Psychology Editor), 6 Abercromby Square, Liverpool 7;
for *The British Journal of Mathematical and Statistical Psychology* should be sent to
Professor R. J. Audley, Department of Psychology, University College London, Gower Street, W.C.1;
for *The British Journal of Educational Psychology* should be sent to
Professor J. D. Nisbet, Department of Education, University of Aberdeen, Old Aberdeen;
for *The Monograph Supplements* should be sent to
Professor I. M. L. Hunter, Department of Psychology, The University, Keele, Staffs., whose decision regarding the publication of all Monograph Supplements will be final. No MS. can be considered for publication in any year unless it is in the hands of the Editor by 1 September of the preceding year.

SEPARATES

Contributors receive on application fifty copies of their papers free. Additional copies of separates may be obtained in units of fifty copies.

SUBSCRIPTIONS TO THE BRITISH PSYCHOLOGICAL SOCIETY

The subscription for Fellows, Associates and Graduate Members is £8 10*s*. (£7 10*s*. if paid by Banker's Order). They are entitled to purchase any or all of the following Journals at the special rates of £1 10*s*. per Journal per annum: *The British Journal of Psychology*, *The British Journal of Medical Psychology*, *The British Journal of Social and Clinical Psychology*, *The British Journal of Mathematical and Statistical Psychology* and *The British Journal of Educational Psychology*.

Information on membership and activities may be obtained from the Secretary, The British Psychological Society, Tavistock House South, Tavistock Square, London, W.C.1.

THE BRITISH JOURNAL OF MATHEMATICAL AND STATISTICAL PSYCHOLOGY

Volume 20, *Part* 2, 1967

CONTENTS

SHALLICE, T.
Temporal Summation and Absolute Brightness Thresholds *page* 129

JONCKHEERE, A. R. and BOWER, GORDON H.
Non-Parametric Trend Tests for Learning Data 163

RODGER, R. S.
Type II Errors and their Decision Basis 187

MCDONALD, RODERICK P.
Factor Interaction in Nonlinear Factor Analysis 205

PHILLIPS, J. P. N.
A Procedure for Determining Slater's *i* and all Nearest Adjoining Orders 217

BAKER, FRANK B.
The Effect of Criterion Score Grouping upon Item Parameter Estimation 227

NOTES AND CORRESPONDENCE
A Note on a Generating Assumption in McGee's Multidimensional Analysis of 'Elastic' Distances DB 239
A Reply to Some Criticisms of Elastic Multidimensional Scaling 243

COMPUTER SECTION
APL: A Computing Language Designed for the User 249

COMPUTER PROGRAM ABSTRACTS 261

BOOK REVIEWS 263

Printed in Great Britain by
Taylor & Francis Ltd., Red Lion Court, Fleet Street, London E.C.4